Structured Product Labeling (SPL)
Implementation Guide with Validation Procedures

Version 1 Revision 202104161249

1 Introduction 3
1.1 Organization 3
1.2 Validation Procedures 3
2 SPL Documents in General 5
2.1 SPL Header 5
2.1.1 General 5
2.1.2 XML references 5
2.1.3 Document information 6
2.1.4 Author Information 8
2.1.5 Identified Organizations 10
2.1.6 Address 10
2.1.7 Telecommunication Addresses 11
2.1.8 Contact Party 13
2.1.9 "Doing Business As" (DBA) Name 14
2.1.10 Core Document Reference 14
2.1.11 Predecessor Document 15
2.2 SPL Body 16
2.2.1 Sections and subsections 16
2.2.2 Text 19
2.2.3 Images 24
2.2.4 Highlights 26
2.2.5 Product Data Elements Section 28
14 Substance Indexing 30
14.1 Header 31
14.1.1 Document type 31
14.1.2 Author information 32
14.2 Body 32
14.2.1 Substance Indexing Section 33
14.2.2 Substance Indexing – Substance Identification 33
14.2.3 Substance Name Detail 34
14.2.4 Substance Mappings – Equivalence 35
14.2.5 [RESERVED] 36
14.2.6 Moieties in General 36
14.2.7 Structural Unit 39
14.2.8 Chemical Structure 42
14.2.9 Stereochemistry 47
14.2.10 Optical Activity 48
14.2.11 Structural Repeat Units of Polymers 48
14.2.12 Irregular Substituents and R-Groups 55
14.2.13 Structural Modifications 59
14.2.14 Monomer Substitutions 60

14.2.15 Monomer Deletion 64
14.2.16 Structural Attachment Modification 66
14.2.17 Markush Structures / R-Group Modification 67
14.2.18 Authority Citation for Organisms 74
14.2.19 Processes 74
14.2.20 Process Interactor 75
14.2.21 Component Process 76
14.2.22 [RESERVED] 77

1 Introduction

Structured Product Labeling (SPL) is a Health Level Seven (HL7) standard based on Clinical Document Architecture and HL7 Reference Information Model (RIM) accredited by the American National Standards Institute (ANSI) for the exchange of product information. Structured Product Labeling documents include a header and body. The header includes information about the document such as the type of product, author and versioning. The body of the document includes product information in both structured text and data element formats. The United States Food and Drug Administration (FDA) uses SPL documents to exchange information covering a growing number of product related topics.

This document provides technical conformance criteria for SPL documents used by FDA. This combines the information previously covered in separate implementation guide and validation procedures documents.[1] A link to the latest SPL schema and controlled terminology used in SPL and other technical documents may be found on the FDA Data Standards Council web site at: https://www.fda.gov/industry/fda-resources-data-standards/structured-product-labeling-resources.

1.1 Organization

This document is divided into three parts. The first part of this document describes the technical conformance criteria that are applicable to header and body of the SPL document Independent of the information being exchanged. The second part of the document describes product related technical conformance criteria. The third part describes the technical conformance criteria applicable to the type of information being exchanged.

1.2 Validation Procedures

Detailed validation procedures are presented at the end of most sub-sections and are clearly marked with the heading "Validation Procedures." These procedures can be used by humans as check-lists to verify if their submission is correct. The validation procedures are written specific and operational so that they may be checked by systems processing SPL documents. Each validation procedure has a unique paragraph number. These paragraph numbers are generally stable over time, but they may change between versions of the document when – rarely – a validation procedure

[1] Instead of 2 documents that both contain details on the structure of SPL files for various purposes with examples, explanations and conformance criteria at varying degree of detailing, the combined document is a systematic compilation of all such technical information in a new topical organization. As SPL is used for an increasing number of different types of products or aspects about products, the old organization became difficult to read and to maintain consistently. The new unified implementation guide with topical organization combines the discussion of consideration and detailed technical conformance rules for each aspect or use of SPL in one place.

is inserted between existing ones; normally, however, new validation procedures are appended to the end of their respective sub-sections.

2 SPL Documents in General

2.1 SPL Header

2.1.1 General

Validation Procedures

2.1.1.1 XML is well formed and valid against the schema

2.1.1.2 There are no data elements and attributes in addition to those described in this document

```
<relatedDocument typeCode="REFR">
   <relatedDocument>
      <setId extension="NCT01352845" root="2.16.840.1.113883.3.1077"/>
   </relatedDocument>
</relatedDocument>There are no spaces in codes
```

2.1.1.4 Codes do not have a codeSystemName attribute

2.1.1.5 Display names are case insensitive

2.1.1.6 There are no spaces in id extensions

2.1.1.7 Letters in Globally Unique Identifiers (GUID) are lower case

2.1.1.8 There are no empty or incomplete elements except, in certain circumstances, code, title, text, and time (an id has a root, a code has a code system).

2.1.1.9 Characteristics have a class code of “OBS” or no class code at all.

2.1.1.10 There is no confidentiality code on anything but inactive ingredients, identified substance, registrant, and assigned establishments outside establishment registrations.

2.1.1.11 If there is a confidentiality code, then the code is “B” and the codeSystem is “2.16.840.1.113883.5.25”

2.1.2 XML references

This information includes the location of the current stylesheet for the FDA view of the SPL and the location of the current schema. The start of the SPL file is the same for every SPL document and is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<?xml-stylesheet
   href="https://www.accessdata.fda.gov/spl/stylesheet/spl.xsl"
   type="text/xsl"?>
<document xmlns="urn:hl7-org:v3"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="urn:hl7-org:v3
                       https://www.accessdata.fda.gov/spl/schema/spl.xsd">
```

Validation Procedures

2.1.2.1 XML reference is for version 1.0 and encoding “UTF-8”.

2.1.2.2 There is an xml-stylesheet reference to “https://www.accessdata.fda.gov/spl/stylesheet/spl.xsl”.

2.1.2.3 The schemaLocation of the urn:hl7-org:v3 namespace is provided as “https://www.accessdata.fda.gov/spl/schema/spl.xsd”.

2.1.2.4 There are no processing instructions other than the xml and xml-stylesheet declarations.

2.1.2.5 There are no comments.

2.1.2.6 SPL file name is the document id followed by “.xml”

2.1.2.7 A submission contains only the SPL file whose name ends in ‘.xml’ and, if appropriate, associated image files whose names end in '.jpg' except if the document types are Wholesale Drug Distributors and Third-Party Logistics Facility Report (75030-7) or REMS document (82351-8) having associated pdf files whose names end in '.pdf'.

2.1.2.8 All image files associated with the SPL document are actually referenced from that SPL document.

2.1.3 Document information

This provides basic information for the identity of the particular document, its type, title, date and versioning as a member of a document set.

Terminology: The SPL document types are from LOINC. This code provides information about the subject matter of the document e.g., prescription animal drug.

```
<document>
  <id root="50606941-3e5d-465c-b4e0-0f5a19eb41d4"/>
  <code code="51725-0" codeSystem="2.16.840.1.113883.6.1"
        displayName="Establishment registration"/>
  <title>Establishment Registration</title>
  <effectiveTime value="20070424"/>
  <setId root="a30accef-f437-4136-808c-9ed4ada5fcf8"/>
  <versionNumber value="1"/>
```

- The <id root> is a Globally Unique Identifier (GUID) and is unique for each version of the document. Letters used in a GUID are lower case.
- The <code> is the LOINC code which provides information on the document type.
- The <title> data element is used for the title for the document, if necessary. Images are not included in the title. Multiple lines may be used in the title with each line separated by the line break
 tag. (note: all titles can also be as follows: <title mediaType="text/x-hl7-title+xml">).
- The <effectiveTime> provides a date reference to the SPL version including the year, month and day as yyyymmdd.
- The <setId> is a GUID and is a unique identifier for the document that remains constant through all versions/revisions of the document.
- The <versionNumber> is an integer greater than zero that provides a sequence to the versions of the document.

Validation Procedures

2.1.3.1 There is a document id

2.1.3.2 id root is a Globally Unique Identifier (GUID).

2.1.3.3 id does not have an extension.

2.1.3.4 id does not match any other id in the document.

2.1.3.5 id (document id) is unique across all documents, sections and any other ids

2.1.3.6 There is a code

2.1.3.7 Code system is 2.16.840.1.113883.6.1

2.1.3.8 Code comes from the *Document type* list

2.1.3.9 Display name matches the code

2.1.3.10 There are no figures in the title.

2.1.3.11 There is an effective time with at least the precision of day in the format YYYYMMDD

2.1.3.12 There is a set id

2.1.3.13 The set id is a GUID

2.1.3.14 There is a version number

2.1.3.15 Value of version number is a whole number > 0

2.1.3.16 Value of version number is greater than the value of any previously submitted version for the same set id

2.1.3.17 The preceding version of this set id has not been replaced by a document with a different set id, i.e., this set id has not been referenced as a related document of type "replace" (RPLC) from a document with a different set id.

2.1.4 Author Information

The author information is represented as follows:

```
<document>
  <author>
    <assignedEntity>
      <representedOrganization>
```

Many times the author information is used to represent details on the businesses responsible for the products. This includes the labeler and registrant and establishments involved in manufacturing:

```
<author>
  <assignedEntity>
    <representedOrganization><!-- labeler -->
      <assignedEntity>
        <assignedOrganization> <!-- registrant -->
          <assignedEntity>
            <assignedOrganization> <!-- establishment -->
              <assignedEntity>
                <assignedOrganization><!-- US agent and importers -->
```

The following is a representative coding of the common structures in the header:

```
<document>
  <author>
    <time/>
    <assignedEntity>
      <representedOrganization><!-- labeler -->
        <id extension="DUNS Number" root="1.3.6.1.4.1.519.1"/>
        <id extension="NDC Labeler Code" root=”2.16.840.1.113883.6.69”/>
        <name>business name</name>

        <contactParty>
          <addr>
            <streetAddressLine>address</streetAddressLine>
            <city>city</city>
            <state>state</state>
            <postalCode>postal code</postalCode>
            <country code="country code"
                     codeSystem="1.0.3166.1.2.3">country name</country>
          </addr>
          <telecom value="tel:telephone number"/>
          <telecom value="mailto:email address"/>
          <contactPerson>
            <name>contact person name for labeler</name>
          </contactPerson>
        </contactParty>

        <assignedEntity>
          <assignedOrganization><!-- registrant -->
            <id extension="DUNS number” root="1.3.6.1.4.1.519.1"/>
            <name>business name</name>

            <contactParty><!-- same structure as above --></contactParty>

            <assignedEntity>
              <assignedOrganization><!-- establishment -->
                <id extension="DUNS number" root="1.3.6.1.4.1.519.1"/>
                <id extension="FDA establishment identifier"
                    root=”2.16.840.1.113883.4.82”/>
                <name>Establishment name</name>
                <addr><!-- as above --></addr>
                <contactParty><!-- as above --></contactParty>

                <assignedEntity>
                  <assignedOrganization><!-- U.S. agent -->
                    <id extension="DUNS Number" root="1.3.6.1.4.1.519.1"/>
                    <name>business name</name>
                    <telecom value="tel: telephone number"/>
                    <telecom value="mailto: email address"/>
                  </assignedOrganization>

                  <performance>
                    <actDefinition>
                      <code code="C73330"
                            codeSystem="2.16.840.1.113883.3.26.1.1"
                            displayName="display name"/>
                    </actDefinition>
                  </performance>
                </assignedEntity>
              </assignedOrganization>
```

```
              <performance>
                <actDefinition>
                  <code code="establishment business operation code"
                        codeSystem="2.16.840.1.113883.3.26.1.1"
                        displayName="display name"/>
                </actDefinition>
              </performance>
            </assignedEntity>
          </assignedOrganization>
        </assignedEntity>
      </representedOrganization>
    </assignedEntity>
  </author>
</document>
```

2.1.5 Identified Organizations

Most organizations are identified using Dun and Bradstreet identifiers (DUNS numbers). These are identifiers with the root 1.3.6.1.4.1.519.1 and an extension.

```
<representedOrganization>
  <id extension="DUNS Number" root="1.3.6.1.4.1.519.1"/>
```

The only reason for an organization not being identified is if the organization remains anonymous but has sub-organizations (e.g., a listing file may not contain any registrant information)

```
<representedOrganization>
  <id extension="DUNS Number" root="1.3.6.1.4.1.519.1"/>
  <name>buiness name</name>
  <assignedEntity>
    <assignedOrganization>
      <!-- pass-through organization without ids or name -->
      <assignedEntity>
        <assignedOrganization>
          <id extension="DUNS Number" root="1.3.6.1.4.1.519.1"/>
          <name>buiness name</name>
```

Validation Procedures

2.1.5.1 One id is a DUNS number with the root 1.3.6.1.4.1.519.1

2.1.5.2 The id (DUNS number) with the root 1.3.6.1.4.1.519.1 has a 9-digit extension

2.1.5.3 There is a name, except if the document is an FDA-Initiated Compliance Action for a Establishment Registration (89600-1).

2.1.6 Address

For addresses (addr) the following rules apply

```
<addr>
   <streetAddressLine>1625 29th street</streetAddressLine>
   <city>Camden</city>
   <state>NJ</state> <postalCode>08101</postalCode>
   <country code="USA" codeSystem="1.0.3166.1.2.3">USA</country>
</addr>
```

Validation Procedures

2.1.6.1 An address has one or two street address line, city, and country.

2.1.6.2 If there is a country code, then it is an ISO 3-letter country code (code system "1.0.3166.1.2.3").

2.1.6.3 If there is no code attribute, then the country name may be the code, otherwise country is a full country name matching the code.

2.1.6.4 If the country is "USA", then the contact party's address has a state (2-letter abbreviation) and postal code

2.1.6.5 If the country is "USA", then the postal code is 5 digits with optionally a dash followed by 4 numbers

2.1.6.6 If the country is **not** in the *postal code validation* list, then there is a postal code

2.1.7 Telecommunication Addresses

Some elements may have telecommunication addresses. If an element has telecommunication addresses it usually allows for a telephone number and an email address.

```
<contactParty>
   ...
   <telecom value="tel:+1-800-555-1213;ext=112"/>
   <telecom value="mailto:Bob.Jones@acme.com"/>
   ...
</contactParty>
```

However, there are exceptions noted in the validation procedures.

Telecommunication addresses are usually provided for an organization's contact party, in which case telephone and email are the common standard. In no case is telephone or email missing for contact party. But in some cases a 3rd telecommunication address can be provided with the FAX number:

```
<contactParty>
  ...
  <telecom value="tel:+1-800-555-1213;ext=112"/>
  <telecom value="mailto:Bob.Jones@acme.com"/>
  <telecom value="fax:+1-302-123-5433"/>
  ...
</contactParty>
```

Normally telecommunication addresses are associated with specific contact parties of organizations as shown above, such as, for registrant contact party or establishment contact party or labeler contact party. However, in some cases, telephone numbers of physical facilities can be specified directly without contact party:

```
<assignedOrganization>
  ...
  <telecom value="tel:+1-800-555-1213;ext=112"/>
</assignedOrganization>
```

In several Establishment/Facility Registrations/Reporting use cases, traditionally the US Agents and Importers have been provided as abbreviated organizations without the added complexity of a contact party:

```
<assignedOrganization>
  ...
  <telecom value="tel:+1-800-555-1213;ext=112"/>
  <telecom value="mailto:Bob.Jones@acme.com"/>
</assignedOrganization>
```

Validation Procedures

2.1.7.1 There are two <telecom> elements, except if the document type is *Lot Distribution Data* (66105-8) or *Indexing - Substance* (64124-1) there is one telecom element or *Generic Drug Facility Identification* (72090-4 or 71743-9) there may be a third telecom element.

2.1.7.2 One telecom value begins with “tel:” and is a telephone number, except if the document code is *Lot Distribution Data* (66105-8) or *Indexing - Substance* (64124-1).

2.1.7.3 For telephone numbers, the following general rules apply:

2.1.7.4 Telephone numbers are global telephone numbers;

2.1.7.5 Telephone numbers contain no letters or spaces;

2.1.7.6 Telephone numbers begin with “+”;

2.1.7.7 Telephone numbers include hyphens to separate the country code, area codes and subscriber number;

2.1.7.8 US telephone numbers have the format +1-aaa-bbb-cccc where “aaa” is the area code, and “bbb-cccc” the usual digit grouping of a local phone number.

2.1.7.9 Telephone numbers have any extensions separated by “;ext=” (see Uniform Resource Identifier (URI) for Telephone Numbers RFC 3966).

2.1.7.10 If there is a semicolon in the telephone number, then it is followed by ext.

2.1.7.11 One telecom value begins with “mailto:” and encodes an email address.

2.1.7.12 an email address is of the simple form <username>@<dns-name>

2.1.7.13 If there is a third telecom element (fax number), then its value begins with “fax:” and its format is the same as for a telephone number.

2.1.8 Contact Party

For most organizations, a contact party may be specified with a contact person as in the following example:

```
<contactParty>
  <addr>
    <streetAddressLine>1625 29th street</streetAddressLine>
    <city>Camden</city>
    <state>NJ</state> <postalCode>08101</postalCode>
    <country code="USA" codeSystem="1.0.3166.1.2.3">USA</country>
  </addr>
  <telecom value="tel:+1-800-555-1213;ext=112"/>
  <telecom value="mailto:Bob.Jones@acme.com"/>
  <contactPerson>
    <name>Bob Jones</name>
  </contactPerson>
</contactParty>
```

Validation Procedures

2.1.8.1 The contactParty has an address element (addr), except if the document is *Lot Distribution Data* (66105-8), *Wholesale Drug Distributor and Third-Party Logistics Facility Report* (75030-7), *Withdrawal of Wholesale Drug Distributors and Third-Party Logistics Facility Report* (77573-4), *Blanket No Changes Certification Of Product Listing* (86445-4), or *Indexing - Substance* (64124-1).

2.1.8.2 The contactParty has telephone number and email addresses.

2.1.8.3 There is one contact person name.

2.1.9 “Doing Business As” (DBA) Name

```
<assignedOrganization> <!-- facility -->
  ...
  <asNamedEntity> <!-- other “doing business as” name -->
    <code code="C117113" displayName="doing business as"
          codeSystem="2.16.840.1.113883.3.26.1.1"/>
    <name>A.C.M.E. Logistic</name>
  </asNamedEntity>
  <contactParty .../>
```

Validation Procedures

2.1.9.1 There is no “doing business as” (DBA) name element, except if the document type is *Wholesale Drug Distributors and Third-Party Logistics Facility Report* (75030-7).

2.1.9.2 DBA name has a name element

2.1.10 Core Document Reference

For some SPL documents it is permitted to specify a “core document” reference. A document with a core document reference “inherits” all the sections from the referenced core document and may override certain top-level sections with its own sections. A core document reference is specified as follows:

```
<document>
  ...
  <author .../>
  <relatedDocument typeCode="APND">
    <relatedDocument>
      <setId root="20d9b74e-e3d8-4511-9df9-cec2087372fc"/>
      <versionNumber value="1"/>
    </relatedDocument>
  </relatedDocument>
  <component .../>
</document>
```

The reference contains the setId of the referenced core-document. The document and the core-document can develop Independently. The core-document may be updated, but the reference remains to the latest core-document with the same setId. The version number in the reference may be provided to indicate which version of the core-document was used when at the time the referencing document was created or modified.

Validation Procedures

2.1.10.1 There is no document id

2.1.10.2 There is a set id

2.1.10.3 Set id is a GUID

2.1.10.4 Document set id is the set id of a core-document.

2.1.10.5 If there is a version number, then it is a whole number > 0.

2.1.10.6 If there is a version number, then it is less or equal than the version of the current core document with that set id.

2.1.11 Predecessor Document

Other documents may be merged into this document by providing a reference to the other predecessor documents that are replaced by this document. Do not provide a reference to the predecessor document under the same set id as the document being submitted, as this is implicitly given by the set id and incremented version number of this document. Only provide references to documents of different set ids. The reference contains only the id of the other predecessor document, code, the setId and the version number. All these ids must match the ids of the other documents that had previously been submitted.

```
<document>
  ...
  <author .../>
  <relatedDocument typeCode="RPLC">
    <relatedDocument>
      <id root="464239de-45c7-4d2f-a89a-45d303f428bd"/>
      <code code="Other Registration Document Type Code"
            codeSystem="2.16.840.1.113883.6.1"
            displayName="Other Registration Document Type Name"/>
      <setId root="9ea75e1e-84ef-4605-89ff-dd08a4c94f40"/>
      <versionNumber value="3"/>
    </relatedDocument>
  </relatedDocument>
  <component .../>
</document>
```

Validation Procedures

2.1.11.1 There is an id (document id)

2.1.11.2 The id (document id) is a GUID

2.1.11.3 There is a set id

2.1.11.4 The set id is a GUID

2.1.11.5 The set id is different from the present document's set id.

2.1.11.6 There is a version number, which is a whole number > 0.

2.1.11.7 The set id has been previously submitted.

2.1.11.8 Document id and version number match the latest document previously submitted under that set id.

2.1.11.9 Document type matches the latest document type previously submitted under that set id.

2.1.11.10 The referenced document has not already been replaced by another document.

2.2 SPL Body

The body of the SPL document includes structured text such as product labeling and specific data elements such as ingredients.

```
<document> <!-- SPL header material -->
  <component>
    <structuredBody> <!-- SPL bodyt material -->
      <component>
        <section>
```

2.2.1 Sections and subsections

```
<component>
   <section>
      <id root="62abedf9-6bde-4787-beb0-abd214307427"/>
      <code  code="34067-9"
             codeSystem="2.16.840.1.113883.6.1"
             displayName="Indications and Usage"/>
      <title>Indications and Usage</title>
      <text>labeling text</text>
      <effectiveTime value="20070822"/>
      <component/>
```

Sections and subsections have id, title, and code. LOINC codes are used for sections and subsections codes.

The <title>, if necessary, of the sections and subsections and order of the sections and subsections in the SPL are used to render the labeling contents. The numbering for the sections and subsections are included in the <title> text.

In the SPL schema, the <structuredBody> element contains multiple <component>s, and each <component> contains a <section>.

Sections are used to aggregate paragraphs into logical groupings. The order in which sections appear in an SPL document is the order the sections will appear when displayed (rendered) using the standard stylesheet. Major sections defined by the appropriate labeling regulations (e.g., 21 CRF 201.56 and 57 for human prescription drugs and 201.66 for human over the counter drugs) such as Indications and Usage are assigned LOINC codes. Sections that have not been assigned a LOINC code are assigned the LOINC code for "SPL Unclassified Section". Major sections may also be defined by parts of a container or carton label (e.g., Principal Display panel).

```
<section>
  <!-- this section's id, codes -->
  <text>
    <!-- actual text content in "narrative block" markup -->
  </text>
```

Each section has a unique identifier (<id>), an <effectiveTime>, and a LOINC code (i.e., the <code> element). A section may or may not contain a <title>.

The human readable content of labeling is contained within the <text> element in the <section>. The <section> can be nested to form sub-sections. The schema for subsections in SPL requires that the nested <section> tag first be nested inside a <component> tag. Use nested sections to relate paragraphs. The section tag applies to all of the nested sections. By nesting sections, computer systems can use the section tags in SPL to display information useful for the care of patients. If information is not associated with the tag, it will not be displayed.

```
<section>
  <!-- this section's id, codes -->
  <text>
    <!-- actual text content in "narrative block" markup -->
  </text>
  <component>
    <section>
      <!-- subsection content -->
    </section>
  </component>
  <component>
    <section>
      <!-- subsection content -->
    </section>
  </component>
</section>
```

Using the following principles for markup of text information improves access to information in labeling:

- Capture the section heading using the <title> element rather than placing the text of the title within the <text> element. This allows computer systems to use and display this information properly.

- Capture the section heading even when the printed label does not include a heading. For example, tagging a pregnancy statement as a section in a label that does not have a heading for pregnancy is useful. Computer systems will be able to use the tag to capture the pregnancy use statement. Omitting the <title> would prevent the heading from appearing when the SPL is rendered.

- Link different parts of the labeling using the ID attribute to the <section> element. For example, <section ID="Clin_Pharm_Section"> serves as the target of a <linkHtml> element. Linking to the ID attribute of a section allows the link to 'reference' the section entirely, e.g., for retrieval of a whole section in a non-browser interface.

- For container or carton labels, when capturing text and figures outside the Drug Facts or equivalent sections, separate the text and figures for each concept using <paragraph> tags.

- The order of the placement of information is the content of the package insert, the content of the patient information and the carton and container labels with images.

Validation Procedures

2.2.1.1 Each section has zero to many subsections

2.2.1.2 Each section and subsection has an id root and no extension

2.2.1.3 id root (section id) is a GUID

2.2.1.4 id does not match any other id in the document

2.2.1.5 id (section id) does not match any other id across all sections, documents, or any id other than the id of the same section previously submitted

2.2.1.6 Each section and subsection has a code

2.2.1.7 Code system is 2.16.840.1.113883.6.1

2.2.1.8 Display name matches the code

2.2.1.9 Each section has an effective time with at least the precision of day in the format YYYYMMDD, except the *Product Data Elements* section (48780-1) of *Lot Distribution Data* (66105-8), *Human Compounded Drug Reporting* (75031-5), *Animal Compounded Drug Label* (77647-6) and *Indexing - Warning Letter Alert* (77288-9) documents, and the Indexing Data Elements section (48779-3) of *Indexing - Biologic or Drug Substance* (77648-4) and *Indexing - Warning Letter Alert* (77288-9).

2.2.1.10 There are no figures in the title for a section or subsection.

2.2.1.11 The section for Medication Guide (42231-1) and Patient Package Insert (42230-3) is not a subsection.

2.2.2 Text

```
<section>
  <text>
    <paragraph>Lorem ipsum dolor sit amet, consectetur adipisicing elit,
sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim
ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip
ex ea commodo consequat. Duis aute irure dolor in reprehenderit in
voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint
occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit
anim id est laborum.</paragraph>
    <paragraph>At vero eos et accusamus et iusto odio dignissimos ducimus
qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores
et quas molestias excepturi sint occaecati cupiditate non provident,
similique sunt in culpa qui officia deserunt mollitia animi, id est laborum
et dolorum fuga.</paragraph>
  </text>
</section>
```

The human readable text content of SPL documents is contained within the <text> element. The actual content is contained within a <paragraph>, <table>, and/or <list>. If a section consists only of nested sections, the <text> tag is not included. Elements that can be used within the <text> element to capture the human readable content of SPL include paragraphs (<paragraph>), lists (<list>), tables (<table>) and images (<renderMultimedia>). Elements permitted as children of the <text> element, used as children of the <paragraph> element or within <table> and <list> include superscripts (<sup>), subscripts (<sub>), links (<linkHtml>), line breaks (
), footnotes (<footnote>), footnote references (<footnoteRef>). Images may be included in the content of labeling using the <renderMultiMedia> tag. This tag may be used as a direct child of <text> for 'block' images or as a child of <paragraph> for inline images.

2.2.2.1 Font effects

There are certain aspects of the rendering of SPL that must be specified in the SPL source to insure that the content of labeling is formatted correctly when rendered. For example:

```
<text>
  <paragraph>The next snippet <content styleCode="bold italics">will appear
as bold italics</content> in the rendering.</paragraph>
```

Will be rendered as:

The next snippet ***will appear as bold italics*** in the rendering.

The <content styleCode=””> can also be nested, for example:

```
<text>
  <paragraph>
     <content styleCode="bold italics"> will appear as bold
italics</content>
```

Can also be represented as:

```
<text>
  <paragraph>
     <content styleCode="bold”><content styleCode="italics”> will appear as
bold italics.</content></content>
```

The values for <styleCode> for font effect are bold, italics and underline. To assist people who are visually impaired, the <styleCode=”emphasis”> is used to prompt computer screen reader programs to emphasize text such as text in a box warning. The bold, italics and underline font effects may be used together with each other and the emphasis styleCode. For example, <content styleCode=”bold”><content styleCode=”emphasis”> </content></content> will appear as bold and will be emphasized by the screen reader programs.

A special styleCode is used for recent major changes (see below).

2.2.2.2 Symbols and special characters

Special characters can be included in the text. Superscripts and subscripts are accomplished using the <sup> and <sub> tags. Because the SPL encoding is UTF-8, any Unicode character can be included as is. Unicode references may also be inserted as either &#dddd; where dddd is the Unicode value in decimal notation or &#xdddd; where dddd is the Unicode value in hexadecimal notation. The font used in the standard stylesheet is a Unicode font assuring that most Unicode characters will be rendered correctly if viewed by a browser supporting this font. The only prohibited characters in XML that can not be directly used are less-than “<” (because SPL XML tags begin with it) and ampersand “&” (because XML entity references begin with it). Use of these two symbols must be replaced by the XML entity references <. and & respectively. For example, “<paragraph>The mean for group 1 was < 13. </paragraph>” will render as “The mean for group 1 was <13.” and “D&C Yellow #10” will render as “D&C Yellow #10”.

2.2.2.3 Footnotes

The SPL schema includes a specific footnote element <footnote>. Footnotes are rendered automatically by the standard SPL stylesheet. <footnoteRef> is used to refer to another (usually earlier) footnote. For example, “<footnote ID=”testNote”>This is the footnote content</footnote>” will generate the following footnote at the appropriate end of a section. “[6]This is footnote content”

The <footnoteRef> element with the appropriate IDREF attribute, e.g., <footnoteRef IDREF=”testNote”/> will display the footnote reference in the text corresponding to the footnote with the same ID, e.g., in this example footnote 6.

Footnotes are rendered by the default stylesheet using Arabic numbers (e.g., 1,2 3,). Within tables, footnotes are rendered using footnote marks in the series: * † ‡ § ¶ # ♠ ♥ ♦ ♣, effectively separating numbered footnotes within general text and footnotes within tables. Footnotes within tables are rendered at the bottom of the table.

2.2.2.4 Lists

All lists are marked up using the <list> tag, and each item in a list is marked with an <item> tag. The ‘listType’ attribute identifies the list as ordered (numbered) or unordered (bulleted). The default numbering and bulleting are controlled by the stylesheet.

```
<text>
  <paragraph>Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed
do eiusmod tempor incididunt ut labore et ...</paragraph>
  <list listType="ordered" styleCode="BigRoman">
    <item>Lorem ipsum dolor sit amet,</item>
    <item>consectetur adipisicing slit</item>
  </list>
  <paragraph>At vero eos et accusamus et iusto ...</paragraph>
</text>
```

Lists featuring a standard set of specialized markers (standard specialized lists) can be created using the styleCode attribute with the <list> element. Options available for ordered lists are:

- Arabic (List is ordered using Arabic numerals: 1, 2, 3)
- LittleRoman (List is ordered using little Roman numerals: i, ii, iii)
- BigRoman (List is ordered using big Roman numerals: I, II, III)
- LittleAlpha (List is order using little alpha characters: a, b, c)
- BigAlpha (List is ordered using big alpha characters: A, B, C)

For example: <list listType="ordered" styleCode="LittleRoman">

For unordered lists the following options exist:

- Disc (List bullets are simple solid discs: ●)
- Circle (List bullets are hollow discs: ○)

- Square (List bullets are solid squares: ■)

For example: <list listType=”unordered” styleCode=”Disc”>

In addition to the standard specialized lists, user-defined characters are also permitted as markers by nesting <caption> within the <item> tag. Note that any character, XML entity, or Unicode symbol, may be used in the <caption>, and that the <caption> for each <item> are not restricted to the same character.

For example: <item><caption>*</caption> the asterisk is used as item marker here.<item>

2.2.2.5 Tables

Tables can be created with the full structure (header (e.g., for column names), body (e.g. for the rows of the table) and footer e.g. for table footnotes)). The element <tbody> is required for an SPL table while the elements <thead> and <tfoot> are optional in the SPL schema. The structure will display a standard typographical table with rules between the caption (table title) and head, the head and body, and the body and <tfoot>. If a <tfoot> element is included and footnotes are present in a table, then footnotes are rendered after the existing content of the <tfoot> element.

It is recommended to always start with a standard table (i.e., <thead> and <tbody> elements) and test to see whether the rendering is unambiguous and interpretable. It is important that the table communicate labeling content not that it duplicates the presentation in word processed or typeset versions of the package insert. In the unusual situation where additional formatting is needed, the rule styleCode specified or certain attributes may be used to modify the table.

The rule codes are as follows (note that the control names are case sensitive).

- Rule on left side of cell is Lrule
- Rule on right side of cell is Rrule
- Rule on top of cell is Toprule
- Rule on bottom of cell is Botrule

Note: More than one rule control may be used in a cell, e.g., <td styleCode code=”Botrule Lrule”>Cell content </td>.

Rule control codes should be used only when necessary for the interpretability of the table. Use of these codes may result in overriding the default rules for tables. Rather than setting the rule for each cell, table rules may also be controlled according to

entire rows or columns by use of the styleCode attributes with <col>, <colgroup>, <thead>, <tfoot>, <tbody> and <tr> elements.

To make rowgroups appear with horizontal rules, use the styleCode attribute "Botrule" with the appropriate <tr> element. The Botrule value is rarely needed on the <td> element.

The preferred method for using vertical rules is to define colgroup with styleCode="Lrule" or "Rrule" (or both). Only if this does not yield the desired vertical rule should the Lrule or Rrule code value with styleCode attributes on the <td> or <th> element itself be used. Note: In general, vertical rules should not be used. Good typography for tables means using few vertical rules.

To merge cells vertically and horizontally, the rowspan and colspan attributes should be used on the <td>element.

To determine the width of a table, the width attribute may be used on the <table> element and to determine the width of a table column, the width attribute may be used on the <col> and <colgroup> elements.

For horizontal alignment, the preferred method for aligning cell content within the margins is to use <col align=".. "/> in the <colgroup> element, though this can be used in the <colgroup> element as well. Valid values for align are "left", "center", "right", "justify" (for full justification of contents within the cells), and "char" (for character alignment within the cells). Using the <col align=".. "/> markup ensures that the contents for all cells in the column share the same alignment.

For vertical alignment, the valign attribute can be used within cells. For cases in which the cell alignment must be different from other cells in the column, align is also available as an attribute on the other table elements, including <td>.

Markup for table footnote is rendered in the <tfoot> tag. This element does not need to be included in SPL; the standard stylesheet will include a <tfoot> tag if a <footnote> element is present within either the <thead> or <tbody> sections. A <tfoot> section should be included in SPL only if there is additional information other than footnotes that needs to be rendered in this section.

For table text spacing, in some instances, the use of a "tab" or text indentation is desirable in a given table cell. In an SPL document, this effect is achieved by using the nonbreaking space ( ) as if it were a "tab" space. As the following snippet of XML shows, two nonbreaking spaces were used to offset the word "Male" from the margin: <td> Male</td>. The nonbreaking space can also be used to keep text in a table from breaking inappropriately due to browser resizing.

2.2.2.6 Hypertext links

SPL offers hypertext linking capabilities generally similar to those found in the HTML specification.

Links are specified by the <linkHtml> construct, where the value for the href attribute of <linkHtml> (the target of the link) is the ID attribute value of a <section>, <paragraph>, <table>, <list>, <content>,<renderMultimedia> element. The stylesheet does not support the styleCode attribute of the <linkHtml> element; if a styleCode is needed for a link, this should be coded via the <content> element within the link as with other text.

2.2.2.7 Recent major changes in labeling text

SPL offers a notation to identify recent major changes in the labeling text including table elements <table> and table data <td>. The recent major text is tagged using the <content styleCode="xmChange">. For example,

```
<text>This is an example of text that is not changed.<content
styleCode="xmChange">This is an example of text that is a recent major
change</content>This is an example of changed text that is not considered a
recent major change</text>
```

Validation Procedures

2.2.2.8 Text is enclosed under <paragraph>, <list>, or <table> elements.

2.2.2.9 The number of table data (<td>) elements is identical to the number of column (<col>) elements in each table (<table>).

2.2.3 Images

The SPL schema uses <observationMedia> elements to identify graphic files to be rendered at the locations where they are referenced by <renderMultiMedia> elements in the <section>. In other words, an image in an SPL will be rendered wherever it is referenced by the renderMultimedia markup, no matter where the observationMedia markup appears. The referencedObject attribute of the renderMultiMedia element identifies the corresponding observationMedia instance by means of its ID identifier such as <renderMultiMedia referencedObject="MM1"/>

```
<section>
  <text>
    <paragraph>...</paragraph>
    <renderMultiMedia referencedObject="MM1"/>
    <paragraph>...</paragraph>
  </text>
```

```
  <component>
    <observationMedia ID="MM1">
      <text>descriptive text</text>
      <value xsi:type="ED" mediaType="image/jpeg">
        <reference value="drug-01.jpg"/>
      </value>
    </observationMedia>
  </component>
</section>
```

The <observationMedia> element does not contain the graphics file, but instead points at the file. The <reference> value is the file name. The file name should not include spaces. The observationMedia identifies the graphic media type (i.e., JPEG). In addition, the observationMedia element includes the text description of the image used by screen reader software for visually impaired users. This is included in the <text> child of <observationMedia>. Note also that observationMedia is always contained within a <component> element as illustrated.

For image placement, if an image is a block image (i.e., should appear in its own space), insert the renderMultimedia tag between <paragraph> elements. If an image is inline (i.e., should appear alongside text), insert the renderMultimedia tag in the text of a <paragraph> as appropriate. Inline images are expected to be uncommon and basically represent symbols that cannot be represented by Unicode characters. In addition, <caption> are not applicable for inline images since these are not offset from the surrounding text.

The SPL stylesheet does not perform any resizing graphics or changing the resolution of graphics files. Thus, all images are rendered in the browser as-is, with all characteristics of the actual graphic file itself. To ensure that a graphic will appear as desired, it is very important that the graphic file is edited to a dimension appropriate for its presentation within the browser. If this is not done, the appearance of the graphic may not be consistent with the narrative content reducing the readability of the file. JPEG image file type using appropriate pixels per inch for images for viewing in a browser using the standard stylesheet.

Validation Procedures

2.2.3.1 There is text

2.2.3.2 Value xsi:type is as above

2.2.3.3 Media type is image/jpeg

2.2.3.4 Reference value is the file name for a valid image

2.2.3.5 Size of image file is less than 1 MB

2.2.3.6 File is a JPEG image and the name has the extension “.jpg”

2.2.3.7 Image components are referenced at least once in the text of any section.

2.2.3.8 Image reference in text has an image “observationMedia” element with a matching ID in the same document.

2.2.4 Highlights

The actual Highlights of a rendered SPL are constructed from four sources: “boilerplate” text rendered directly from the stylesheet, information from data elements inserted into the boilerplate text, <title> in the header which includes the drug names, dosage form, route of administration, controlled substance symbol and year of initial US approval, and text blocks corresponding to each major highlights part (Highlights text). Highlights section titles are derived from the FPI section LOINC codes. The Highlights text is captured for the following sections: Microbiology, Boxed Warning, Recent Major Changes, Indications and Usage, Dosage and Administration, Dosage Forms and Strengths, Contraindications, Warnings and Precautions, Adverse Reactions, Drug Interactions and Use in Specific Populations.

The text blocks for Highlights are coded with the <excerpt> <highlight> elements of the major section of labeling in which they are contained.

```
<section>
  <excerpt>
    <highlight>
      <text>...</text>
```

For example, the Highlights for Indications and Usage are located with the Indications and Usage section of the labeling. The Highlights text is placed under the main section and not under subsections. The following is an example:

```
<component>
  <section>
    <id root="47ef84cd-8314-48c3-8ee2-bdff3087f83f"/>
    <code code="43685-7" codeSystem="2.16.840.1.113883.6.1"
          displayName="warnings and precautions section"/>
    <title>5 WARNINGS AND PRECAUTIONS</title>

    <excerpt>
      <highlight>
        <text>
          <list listType="unordered">
            <item>Aplastic anemia has been observed in 8% of recipients and
is irreversible in the majority of patients who experience this. (<linkHtml
href=”#Section_5.1”>5.1</linkHtml>)</item>
            <item>Monitor for hematological adverse reactions every 2 weeks
through the second month of treatment (<linkHtml
href=”#Section_5.2”>5.2</linkHtml>)</item>
          </list>
        </text>
      </highlight>
    </excerpt>
```

```
    <component>
      <section ID="Section_5.1">
        <id root="a857689e-9563-43c0-a244-8a6d5a25966a"/>
        <title>5.1 Aplastic anemia</title>
        <text>
          <paragraph>Aplastic anemia has been observed in…..</paragraph>
        </text>
      </section>
    </component>
  </section>
</component>
```

This example illustrates the following principles:

a. The <text> block for the Highlights is included as the <excerpt> <highlight> <text> children of the respective section. In the example above, the text block rendered in the highlights section is the child of the "Warnings and Precautions" section.

b. The coding of the highlights text block is not in a subsection.

c. The text block is rendered similar to any other text block, although in a location separate from its actual position in the rendered SPL document.

d. Links to the section or subsection where the primary content exists are explicitly entered in the Highlights text block.

e. Section numbering is included in the title of sections and subsections (e.g., '5' and '5.1', above).

Highlights and labeling boilerplate items include:

- Statement -"Highlights of Prescribing Information"
- Highlights section titles
- Patient counseling statement with information taken from FPI section LOINC codes for patient information sections, specifically information for patient section (34076-0), SPL Medguide section (42231-1), SPL patient package insert section (42230-3) and SPL supplemental patient material (38056-8)
- Revision date is taken from the effective time
- Full Prescribing Information: Contents
- Statement – "Full Prescribing Information"

Validation Procedures

2.2.4.1 There may be excerpts (sections with highlights text).

2.2.4.2 Excerpts occur only in sections with the following codes: 34066-1 (Boxed Warning), 43683-2 (Recent Major Changes), 34067-9 (Indications and Usage), 34068-7 (Dosage and Administration), 43678-2 (Dosage Forms and Strengths), 34070-3 (Contraindications), 43685-7 (Warnings and Precautions), 34084-4 (Adverse Reactions), 34073-7 (Drug Interactions), 43684-0 (Use in Specific Populations), 49489-8 (Microbiology)

2.2.4.3 If there is an excerpt, then it only has highlight text.

2.2.4.4 An excerpt in the adverse reactions section (34084-4) includes the statement: "to report suspected adverse reactions" and "1-800-332-1088" (different telephone number for documents of type 53404-0 – “Vaccine Label”).

2.2.4.5 If there are highlights excerpts, then the title for the SPL file includes the text string (without the quotation marks): “These highlights do not include all the information needed to use” “see full prescribing information for” and “Initial U.S. Approval”

2.2.5 Product Data Elements Section

Currently most of the time the product data elements are in a separate section of their own followed by the content of labeling sections that contain only text and no data elements. Product data element section and other special data elements sections are described in Section **Error! Reference source not found.** below; this section describes the features used from the free text (so called “narrative”) part of the SPL documents.

```
<document>          <!-- SPL header material here -->
  <component>
    <structuredBody><!-- SPL body material here -->

      <component>
        <section>   <!-- Product data element section -->
          <code code="48780-1" codeSystem="2.16.840.1.113883.6.1"
                displayName="SPL product data elements section"/>

          <subject>
            <manufacturedProduct>
              <!-- product data elements -->
            </manufacturedProduct>
          </subject>

        </section>
      </component>
                  <!-- Other content of labeling material -->
      <component>
        <!-- ... -->
```

The beginning of the product data elements is as follows

```
<component>
  <section>
    <id root="e13a985b-f706-a5c8-e8ef-73891eb1c697"/>
      <code code="48780-1"
            codeSystem="2.16.840.1.113883.6.1"
            displayName="SPL product data elements section"/>
      <effectiveTime value="20070424"/>
      <subject>
        <manufacturedProduct>
```

Validation Procedures

2.2.5.1 Code, code system and display name are as above

2.2.5.2 There is one or more product, except in *Human Compounded Drug Reporting* (75031-5) and *Animal Compounded Drug Label* (77647-6) documents.

2.2.5.3 There is an effective time with at least the precision of day in the format YYYYMMDD.

2.2.5.4 There is only one product data element section.

14 Substance Indexing

The scope of this specification extends to definitional information only, i.e., the properties and characteristics that distinguish one substance from another; not to supplementary information, such as properties that can be derived (e.g., molar mass, viscosity) or even if not derived from or are dependent on other definitional characteristic. Placement in hierarchical classifications, such as, for botanical or zoological taxonomy, or other generalizations such as “organic” vs. “inorganic” substance, etc. are not definitional in nature and change over time, and are therefore excluded. Names and codes are supported but not required as this introduces ambiguity and contention: while people might agree that a certain chemical structure exists, they may not agree on the correct or preferred name; and different people might have diverging understanding of the substance called by the same name.

The following kinds of substances are currently in scope of this specification:

- Substances defined by their chemical structure
  - Chemical compounds commonly referred to as “small molecules”
  - Non-stoichiometric chemical compounds
  - Template-driven biopolymers such as proteins and
  - nucleic acids.
  - Stochastic polymers
  - Structurally defined substances with variants defined by Markush R-group attachments.
- Substances that are extracted or otherwise derived from natural source material and whose chemichal structure is not suffiently understood to defined them structurally. Often these natural extracts are an unknown mixture of related but diverse substances, hence also called “structurally diverse”.
- Mixtures including hybrids from the above categories.

The purpose of this specification is to provide for unique canonical definitions of every substance, i.e., variations of alternative ways to specify the same substance shall be minimized as much as possible.

Overview of this Specification

In Sections 14.1 Header and 14.2.1 Substance Indexing Section.5 the general document features for substances and some non-definitional data elements such as names and codes are specified. This must be understood in conjunction with Section 1 Introduction and Section 2 SPL Documents in General.

Section **Error! Reference source not found.** through .17 specify data elements for defining substances by their chemical structure. The specific approach to specifying

substances is structurally is based on the concept of moieties introduced in Section **Error! Reference source not found. Error! Reference source not found.**.

The most fundamental kind of moiety is the Structural Unit (Section 14.2.7) which holds some representation of chemical structure. The structural representation for small molecules, template-driven biopolymers, or stochastic polymers are described in Sections 14.2.8 through .12

The special type of moiety is the Structural Modification (see Section 14.2.13 through .17) which is used to described irregularities or variability of the corresponding Structural Units. In biopolymers and stochastic polymers these modifications are *substitutions* (Section 14.2.14) and in biopolymers also deletions (Section 14.2.15). Only when there is indeterminism, the Structural Attachment Modification (Section 14.2.16) and the Markush Structures / R-Group Modification (Section 14.2.17) are available to represent such classes of inderterministic variant structures.

Sections 14.2.18 through .22 specify data elements for substances that cannot be described by their chemical structure and are derived from natural source products (plants, animals), those are defined primarily with an Authority Citation for Organisms (Section 14.2.18) and then minimal derivation Processes (Section 14.2.19) by which the substance is produced. While these processes reflect manufacturing processes, e.g., clipping of leaves and boiling them to make tea, or drying such extract to make extract powder, or picking fruit and pressing it to derive juice, the point is not to specify the manufacturing process in detail, only enough to define the substance which many detailed manufacturing processes could produce. As necessary reagents, solvents, and catalysist can be specified as Process Interactor (Section 14.2.20) and other details.

14.1 Header

14.1.1 Document type

```
<document>
   <code code="64124-1" codeSystem="2.16.840.1.113883.6.1"
         displayName="Indexing - Substance"/>
```

Validation Procedures

14.1.1.1 Document code is as above

14.1.1.2 If a document with the same set id has been previously submitted, then it is of the same type.

14.1.1.3 The prior version of this document has not been replaced with a document of a different setId.

14.1.2 Author information

The official substance indexing files are maintained by FDA:

```
<document ...>
  <!-- ... -->
<author>
  <time/>
  <assignedEntity>
    <representedOrganization>
      <id root="1.3.6.1.4.1.519.1" extension="927645523"/>
      <name>Food and Drug Administration</name>
```

Other entities may submit substance indexing files and identify themselves by a DUNS number and name along with a contact person with email:

```
<document ...>
  <!-- ... -->
<author>
  <time/>
  <assignedEntity>
    <representedOrganization>
      <id root="1.3.6.1.4.1.519.1" extension="123456780"/>
      <name>Advanced Drug Development Corp.</name>
      <contactParty>
        <telecom value="mailto:fred.frederickson@advanceddrx.com"/>
        <contactPerson>
          <name>Fred Frederickson</name>
        </contactPerson>
      </contactParty>
```

Validation Procedures

14.1.2.1 There is an author.

14.1.2.2 Author organization of official substance indexing files is FDA

14.1.2.3 Other submitters include their DUNS number and name with the author organization

14.1.2.4 If the author organization is not FDA or EPA, then there is a contactParty.

14.2 Body

```
<section>
  <id root="ffabedf9-6bde-4787-beb0-abd214307427"/>
  <code code="48779-3" codeSystem="2.16.840.1.113883.6.1"
        displayName="SPL Indexing Data Elements Section"/>
    <title/>
    <text/>
    <effectiveTime value="20101007"/>
    <subject>
```

Validation Procedures

14.2.1 Substance Indexing Section

14.2.1.1 If the document type is 64124-1, then the document contains one SPL Indexing Data Elements section as above.

14.2.1.2 Value of effective time is same as value of effective time in document information.

14.2.2 Substance Indexing – Substance Identification

There are one or more substance, the first being the main substance to be defined in this document and complying with the validation procedures in this and the following sections. The substances after the first one, are locally defined and used for the definition of the main one.

Main substance:

```
<section>
  <subject>
<identifiedSubstance>
  <id extension="P88XT4IS4D" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance>
    <code code="P88XT4IS4D" codeSystem="2.16.840.1.113883.4.9"/>
    ...
```

Followed by auxiliary substances (if any):

```
<section>
  <subject>
    <identifiedSubstance ... main substance .../>
  </subject>
  <subject>
    <identifiedSubstance><!-- auxiliary substance -->
      <id extension="Local Code String" root="Document Id (UUID)"/>
      <identifiedSubstance>
        <code code="Base Substance UNII"
              codeSystem="2.16.840.1.113883.4.9"/>
        ...
```

Substances that are not public, are marked as confidential as follows:

```
<section>
  <subject>
<identifiedSubstance>
  <id extension="P88XT4IS4D" root="2.16.840.1.113883.4.9"/>
  <confidentialityCode code="B" codeSystem="2.16.840.1.113883.5.25"/>
  <identifiedSubstance>
    <code code="P88XT4IS4D" codeSystem="2.16.840.1.113883.4.9"/>
    ...
```

Validation Procedures

14.2.2.1 There is one or more substances, the first being the main substance to be defined in this document, the following substances being locally defined and used for the definition of the main one.

14.2.2.2 There is one substance code, except for the main to be defined substance if the author is not FDA.

14.2.2.3 If this is the main to be defined substance and the author is FDA, then the code system is 2.16.840.1.113883.4.9 (UNII).

14.2.2.4 If this is the main to be defined substance and the author is EPA, then the code system is 2.16.840.1.113883.6.275.

14.2.2.5 If the author is *not* FDA, then the code system is not 2.16.840.1.113883.4.9 (UNII).

14.2.2.6 If this is an auxiliary locally defined substance, then the code system is the same as the document id root.

14.2.2.7 The parent element's id root and extension are the same as this element's codeSystem and code, respectively.

14.2.2.8 There may be one substance name.

14.2.2.9 If a substance indexing file with the same code has previously been submitted, then the name is not different.

14.2.2.10 The same main to be defined substance is not described in a substance indexing document with a different set id.

14.2.2.11 There is no document with the same set id but a different main to be defined substance.

14.2.3 Substance Name Detail

Every name, including the primary name, is described in detail in the substance name detail structure:

```
<identifiedSubstance>
  <identifiedSubstance>
    <asNamedEntity>
      <code code="C43707" displayName="primary name"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>
      <name>alizapride</name>
```

Validation Procedures

14.2.3.1 If there is a name, then there is one or more name detail elements ("asNamedEntity")

14.2.3.2 Name detail has a code with code system 2.16.840.1.113883.3.26.1.1 and code specifying the type of name as either "primary name" (C43707) or "display name" (C43682, sometimes called an alternative "listing name").

14.2.3.3 Name detail has a name element with the name.

14.2.4 Substance Mappings – Equivalence

Equivalence mappings declare that the substance is considered equivalent with another description of the substance in a different system.

```
<identifiedSubstance>
  <identifiedSubstance>
    <asEquivalentSubstance>
      <definingSubstance>
        <code code="5243513e-5938-b544-afa5-b6aadafe2ad0"
              codeSystem="2.16.840.1.113883.3.2705"/>
      </definingSubstance>
```

Validation Procedures

14.2.4.1 The main to be defined substance may have one or more equivalent substance references.

14.2.4.2 Reference has a code and code system.

14.2.4.3 If the author is FDA, then for the main substance, one of the equivalent substance references is the definition hash code with code system 2.16.840.1.113883.3.2705.

14.2.4.4 Code consists of 32 hexadecimal digits in lower case grouped in 8-4-4-4-12 digits separated by hyphens.

14.2.4.5 For the main substance, definition hash code is not associated with another substance indexing document with a different set id, except if the current document replaces the other document by a related document reference of type *replace* (RPLC) with that other set id.

14.2.4.6 If a prior version of this set id exists, then the definition hash code is the same.

14.2.5 [RESERVED]

14.2.6 Moieties in General

Many structural features of the substances are defined using elements called "moiety". The IUPAC Gold Book, which is the authority of international chemical nomenclature, defines moiety as follows:

> *In physical organic chemistry moiety is generally used to signify part of a molecule, e.g. in an ester R^1COOR^2 the alcohol moiety is R^2O.*[2]

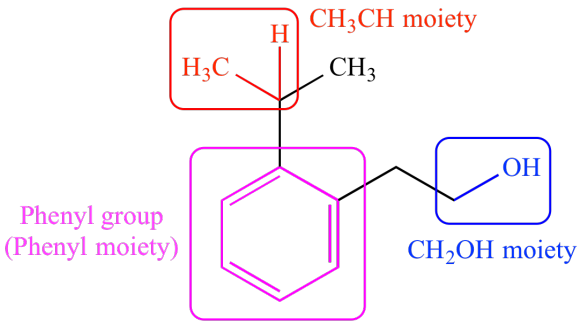


Figure 1: Various Moieties of a Molecule

This can be paraphrased as "any sub-structure of a molecule." The fragment exclusion in this definition is for actual fragment radicals or ions, products of fragmentation. A similar definition in the Illustrated Glossary of Organic Chemistry, Moiety is "*Any part of a molecule. Does not have to be a complete functional group*",[3] so, a moiety can even just be a single atom. In this convention, moieties need not be additive parts of the molecule, and they can overlap. Some moieties are used to additively describe a complex chemical structure, but others may simply delineate features or sites of interests, such as binding sites or other sites with a particular function. Moieties need not be contiguous either, for example, the part of a polymer repeat unit that are chained together are on opposite sides of the molecule, but one may still call them one moiety, because later on one wants to refer to that discontiguous part of the molecule to indicate how they are bound to some other parts of the larger structure.

An example for a moiety element looks as follows:

```
<identifiedSubstance>
  <!-- ... -->
  <identifiedSubstance>
    <!-- ... -->
```

[2] IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). Online version (2019-) created by S. J. Chalk. ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbook; page 1141.
[3] Illustrated Glossary of Organic Chemistry – Moiety. http://www.chem.ucla.edu/~harding/IGOC/M/moiety.html. Retrieved 2020-02-14.

```
    <moiety>
      <code code="Moiety Code" displayName="Moiety Concept Name"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety .../>
    </moiety>
```

The following Table 1 lists all the different kinds of moieties employed by this specification, and what they are used for: Details will be presented in special subsections in the remainder of this chapter that deal with those uses. The types of moieties are:

- "structural unit" – a moiety that additively contributes part of the chemical structure of the entire substance,
- "modification" – a moiety delineating where the base structure is edited in some way, usually by substitution with a substituent
- "connection points" – a moiety of a structure that is used as a substituent, indicating where the substituent connects to the structure that it modifies.

Table 1: Moiety Codes

Name	Code	Type	Used for
Simple Chemical	*no code*	structural unit	all
Protein Subunit	C118424	structural unit	proteins
Polynucleotide	C164242	structural unit	polynucleotides
Polymer	C48803	structural unit	polymers
Mixture Component	C103243	structural unit	all
Structural Repeat Unit	C164243	structural unit	polymers
Structural Modification	C118425	modification	all
Amino Acid Connection Points	C118427	connection points	poteins
Nucleotide Connection Points	C157467	connection points	polynucleotides
Linear SRU Connection Points	C132921	connection points	polymers
Branched SRU Connection Points	C157466	connection points	polymers
Cross-linked SRU Connection Points	C164241	connection points	polymers
R-Group Connection Point	C177645	connection points	Markush groups

Many kinds of moieties have a quantity, with numerator and denominator specifying how much of the part moiety is in the substance.

```
<moiety>
  <quantity>
    <numerator value="1" unit="1"/>
    <denominator value="1" unit="1"/>
  </quantity>
  <partMoiety>
```

The moiety quantity can be by number of parts (unit="1"), or amount of substance (unit="mol"), or – sometimes – by mass (unit="g").

The numerator may be an uncertain range:

```
<moiety>
  <quantity>
    <numerator xsi:type="URG_PQ">
      <low value="1.25" unit="g" inclusive="false" />
      <high value="1.75" unit="g" />
    </numerator>
    <denominator value="1" unit="1"/>
  </quantity>
  <partMoiety>
```

Validation Procedures

14.2.6.1 There may be one or more moieties

14.2.6.2 There is a code, except for simple chemicals.

14.2.6.3 Code system is 2.16.840.1.113883.3.26.1.1

14.2.6.4 Code comes from the moiety code list as shown in Table 1.

14.2.6.5 Display name matches the code.

14.2.6.6 There is a quantity, except if the moiety role code is any of the connection point moieties from Table 1.

14.2.6.7 Quantity has a numerator.

14.2.6.8 If the numerator is an uncertain range, i.e., labeled with an xsi:type attribute of “URG_PQ”, then it has at least one sub-element <low> or <high> for the uncertain range boundaries; with missing low meaning 0, or “at most” the high boundary, and missing high meaning “at least” the low boundary and potentially the maximum possible.

14.2.6.9 If the numerator xsi:type attribute is not specified as “URG_PQ”, then there are no sub-elements low or high.

14.2.6.10 If the numerator xsi:type attribute is not specified as “URG_PQ”, then there are direct value and unit attributes

14.2.6.11 A sub-element low or high has direct value and unit attributes.

14.2.6.12 Numerator has a unit either directly or in the sub-elements low, or high

14.2.6.13 Numerator unit is either mol or g.

14.2.6.14 If the moiety code is *Protein Subunit* (C118424), *Polynucleotide* (C164242), *Polymer* (C48803), *Mixture Component* (C103243) or *Simple Chemical* (no code) then the numerator unit is mol, but not g.

14.2.6.15 If the low or high range boundaries are both present, then both have the same unit.

14.2.6.16 Quantity has a denominator.

14.2.6.17 Denominator has direct value and unit attributes.

14.2.6.18 Denominator has no xsi:type attribute and no low or high range boundary

14.2.6.19 Denominator unit is either mol, or g.

14.2.6.20 If the moiety code is *Protein Subunit* (C118424), *Polynucleotide* (C164242), *Polymer* (C48803), *Mixture Component* (C103243) or *Simple Chemical* (no code) then the denominator unit is mol, but not g.

14.2.6.21 There is a partMoiety element.

14.2.7 Structural Unit

A structural unit moiety has some expression of molecular structure. Substances that are at least partially structurally defined have at least one structural unit, but they can have more than one structural unit. Simple chemicals are defined as one structural unit, but racemates and other enantiomeric mixes are represented as a mixture of each stereo-isomer as its own structural unit. Complex proteins are composed of their different chains, each being its own structural unit. Polymers are composed of one or more “structural repeat units”, just another type of a structural unit.

Moiety “Simple Chemical” is used to define a small molecule with fully defined structure. The structure of this moiety is represented by MOLFILE, SMILES, InChI and InChIKey (see section 14.2.8). InChI is required for unique identification of the structure. Small proteins and nucleic acids (up to 999 atoms and 999 bonds) are also represented as simple chemicals.

Moiety “Mixture Component” is used to define a molecule that is a part of a mixture. The structure of this moiety is represented by MOLFILE, SMILES, InChI and InChIKey (see section 14.2.8). Mixtures may have nested moieties. Combined amount of mixture components should not exceed 1 mol.

Moiety “Protein subunit" is used to define a large polypeptide. Its structure is represented by Amino Acid Letter Sequence (see section 14.2.8).

Moiety “Polynucleotide” is used to define a large polynucleotide. Its structure is represented by Nucleotide Letter Sequence (see section 14.2.8).

Moiety “Polymer” is used to define a stochastic polymer. It must have nested moieties representing structures of fragments also known as “Structural Repeat Units” (see section 14.2.11), as well as amounts of those fragments expressed in mols or grams.

Here is an example of a mixture component:

```
<identifiedSubstance>
  <!-- ... -->
  <identifiedSubstance>
    <!-- ... -->
    <moiety>
      <code code="C103243" displayName="mixture component"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>
      <quantity>
        <numerator value="1" unit="1"/>
        <denominator value="1" unit="1"/>
      </quantity>
      <partMoiety>
        <code code="ABC123XYZ9" codeSystem="2.16.840.1.113883.4.9"/>
      </partMoiety>
      <subjectOf>
        <characteristic>
          <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="Chemical Structure"/>
          ...
        </characteristic>
      </subjectOf>
   </moiety>
   <moiety ... other mixture component .../>
```

A mixture component as shown above has a quantity that represents molar fraction. It is important for calculation of the molar mass of the mixture. Each component adds up to the total not as a whole, but as a fraction (see here for an example.) Often the unit mol will be used, both in numerator and denominator. Since it is widely recognized today that the mol is just a certain big number (Avogadro’s number) of entities, either mol or “1” can be used as the unit. Mass and volume should not be used. For mixtures, numerator should be less than denominator and the sum of mol fractions (numerator/denominator) should equal to 1.

For simple chemical substances that can be defined using only one molecular structure, there would in principle be no need to insert a structural unit as a moiety; instead, one could just add the structure descriptions directly to the substance. However, the advantage of always carrying the structure in a structural unit moiety is that it leads to a uniformity of representation and a direct meaning of the expression “the substance has only a single structural unit”. So, all structurally defined substances have one or more structural units as moieties.

The partMoiety element of a moiety is often empty, because the chemical structure description is connected to the moiety element itself.

```
<identifiedSubstance>
  <!-- ... -->
  <identifiedSubstance>
    <!-- ... -->
    <moiety>
      <code code="C103243" displayName="mixture component"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>
      <quantity>
        <numerator value="0.5" unit="mol"/>
        <denominator value="1" unit="mol"/>
      </quantity>
      <partMoiety/>
```

The partMoiety element of a moiety is *not* empty, however, if the moiety itself is also defined as a substance of its own. This occurs frequently if there is only one moiety which is then in fact the same thing as the entire identified substance. And it also occurs when sub-units are labeled to be referenced later in connecting or binding the various structural units together.

For example, there can be a partMoiety with code and id to declare or refer to a specific chain as subunit 1:

```
<partMoiety>
  <id extension="SU1" root="Document Id" />
  <code code="KAPPA-CHAIN" codeSystem="Document Id" />
```

This is often used to be able to refer to different sub-units of the same structure, such as in composing an antibody from 2 kappa and 2 lambda chains.

```
<identifiedSubstance>
  <id extension="P88XT4IS4D" root="2.16.840.1.113883.4.9"/>
  <identifiedSubstance>
    <code code="P88XT4IS4D" codeSystem="2.16.840.1.113883.4.9"/>
    <name>paclitaxel</name>
    <moiety>
      <quantity>
        <numerator value="1" unit="mol"/>
        <denominator value="1" unit="mol"/>
      </quantity>
      <partMoiety>
        <code code="P88XT4IS4D" codeSystem="2.16.840.1.113883.4.9"/>
      </partMoiety>
```

Note in the above example of a substance that is a simple chemical with only one structural unit, that structural unit moiety has no moiety code of “mixture component” or anything else.

Moieties may also occur nested in other moieties, such as in mixtures of mixtures. For example, in mixtures of different molecules, some of which being in turn enantiomeric mixes.

```
<moiety>
  <partMoiety>
    <moiety>
```

Validation Procedures

14.2.7.1 There may be one or more structural unit moieties

14.2.7.2 If the partMoiety element has an id for document internal references, then the root is equal to the document id root.

14.2.7.3 If the partMoiety element has a code and any modifying properties (e.g. bonds), then there is an id differentiating this moiety from its kind specified by the code.

14.2.7.4 There may be a UNII code (code system 2.16.840.1.113883.4.9).

14.2.7.5 The partMoiety element can have a code referencing an auxiliary substance, locally defined in this document.

14.2.7.6 Code system of locally defined codes is the document id root.

14.2.7.7 Locally defined code is defined as the id of an auxiliary substance in this document.

14.2.7.8 There is no name.

14.2.7.9 Structural units have no other characteristics than *Chemical Structure* (C103240),.*Stereochemistry* (C18188).

14.2.7.10 If the moiety role code is *Polymer* (C48803), then there is one or more nested moieties with role code *Structural Repeat Unit* (C164243).

14.2.8 Chemical Structure

A chemical structure is described as a characteristic of the structural unit and represented as InChI and additionally optionally MOLFILE and SMILES, or alternatively Amino Acid Sequence or Nucleic Acid Sequence.

InChI is required for basic chemical structures because there is only a single InChI string for each structure for simple molecules, ions, and fragments.

MOLFILE representations of structures should be included if available to convey geometry of the molecules and fragments. MOLFILEs must be canonicalized by sorting the atom list in the same order as the InChI atom list.

SMILES may be included, but there is no information in SMILES that is not already in an InChI. SMILES do not have to be canonical SMILES.

The MOLFILE V2000 maximum of 999 atoms is used to decide when to break larger molecules into multiple structural units. This typically happens with sequences. When structures are not determined, such as with polymers and Markush group attachments, then smaller structural units must be used because of that indeterminate nature of the structure.

```
<identifiedSubstance>
  <identifiedSubstance>
    <moiety>
      <partMoiety><!-- Moiety Details --></partMoiety>
      <subjectOf>
        <characteristic>
          <code code="C103240" displayName="Chemical Structure"
                codeSystem="2.16.840.1.113883.3.26.1.1"/>
          <value xsi:type="ED" mediaType="Structural Representation Type">
            <!-- Structural Representation -->
          </value>
```

Known structural representation types are:

Table 2: Structure Representation Type

Representation Type	MIME Media Type
InChI	application/x-inchi
InChIKey	application/x-inchi-key
MOLFILE	application/x-mdl-molfile
SMILES	application/x-smiles
Amino Acid Letter Sequence	application/x-aa-seq
Nucleic Acid Sequence	application/x-na-seq

For example, MOLFILE data is conveyed as follows:

```
<moiety>
      <partMoiety><!-- Moiety Details --></partMoiety>
      <subjectOf>
        <characteristic>
          <code code="C103240" displayName="Chemical Structure"
                codeSystem="2.16.840.1.113883.3.26.1.1"/>
          <value xsi:type="ED"
                 mediaType="application/x-mdl-molfile"
><![CDATA[
  -FDASRS-            2D

38 38  0     1  0  0  0  0  0999 V2000
    4.2690    1.2220    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350   -3.2780    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.5369   -1.7780    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.0010    2.2220    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    6.8671    2.7220    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.0010    1.2220    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.1350    0.7220    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
  ...
  1  7  1  0  0  0  0
  1 33  1  0  0  0  0
  2 15  1  0  0  0  0
  2 37  1  0  0  0  0
  3 17  1  0  0  0  0
  ...
M  END]]></value>
```

InChI representation is conveyed as:

```
    <moiety>
      <partMoiety><!-- Moiety Details --></partMoiety>
      <subjectOf>
        <characteristic>
          <code code="C103240" displayName="Chemical Structure"
                codeSystem="2.16.840.1.113883.3.26.1.1"/>
          <value xsi:type="ED"
                  mediaType="application/x-inchi"
             >InChI=1S/C18H15Cl3N2O.HNO3/c19-14-3-1-13(2-4-14)11-24-18(10-
23-8-7-22-12-23)16-6-5-15(20)9-17(16)21;2-1(3)4/h1-9,12,18H,10-
11H2;(H,2,3,4)/t18-;/m0./s1</value>
        </characteristic>
      </subjectOf>
```

and InChI Key:

```
    <moiety>
      <partMoiety><!-- Moiety Details --></partMoiety>
      <subjectOf>j
        <characteristic>
          <code code="C103240" displayName="Chemical Structure"
                codeSystem="2.16.840.1.113883.3.26.1.1"/>
          <value xsi:type="ED" mediaType="application/x-inchi-key"
            >DDXORDQKGIZAME-FERBBOLQSA-N</value>
        </characteristic>
      </subjectOf>
```

An alternative, more compact way of describing chemical structure, where appropriate, is the Amino Acid Sequence.

```
<identifiedSubstance>
  <identifiedSubstance>
    <moiety>
      <code code="C118424" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="PROTEIN SUBUNIT"/>
      <partMoiety><!-- Moiety Details --></partMoiety>
      <subjectOf>
        <characteristic>
          <code code="C13187" codeSystem="2.16.840.1.113883.3.26.1.1"
                displayName="AMINO ACID SEQUENCE"/>
          <value xsi:type="ED" mediaType="application/x-aa-seq"
            >GFSSLFKAGA KYLLKSVGKA GAQQLACKAA NNCA</value>
```

Amino acid sequences are composed of letters from Table 3, where the capital letters are for the normal L-configuration and the same letters in lower case for the D-configuration. The letter X stands for a non-standard amino-acid. Whitespace characters may be used for visual grouping and have no other significance.

Table 3: L-Amino Acid Letter Codes (D-configuration is lower case)

Code	Amino acid	Code	Amino acid
A	alanine	K	lysine
R	arginine	M	methionine
N	asparagine	F	phenylalanine
D	aspartic acid	P	proline
C	cysteine	S	serine
E	glutamic acid	T	threonine
Q	glutamine	W	tryptophan
G	glycine	Y	tyrosine
H	histidine	V	valine
I	isoleucine		
L	leucine	X	*non-standard amino-acid*

Nucleotide sequences (DNA and RNA sequences) are specified with the value media type “application/x-na-seq”, composed from the letters in Table 4, where capitals stand for DNA and lower case for RNA. The letter X stands for a non-standard nucleotide. Any whitespace character may be used for visual grouping and has no other significance.

Table 4: Nucleotide Letter Codes

Code	Nucleotide
A	deoxyadenosine monophosphate
C	deoxycytidine monophosphate
G	deoxyguanosine monophosphate
T	deoxythymidine monophosphate
U	deoxyuridine monophosphate

a	adenosine monophosphate
c	cytidine monophosphate
g	guanosine monophosphate
t	thymidine monophosphate
u	uridine monophosphate
X	*non-standard nucleotide*

Validation Procedures

14.2.8.1 There are one or more structure representations or sub-moieties, except if the moiety type is *structural modification* (C118425) or *Structural Repeat Unit* (C164243).

14.2.8.2 Code for structure representation is C103240 and code system is 2.16.840.1.113883.3.26.1.1.

14.2.8.3 There is a value is of type ED with media type.

14.2.8.4 Media type is in the list from Table 2: Structure Representation Type above.

14.2.8.5 Value contains text formatted according to the declared structure representation.

14.2.8.6 If value media type is “application/x-aa-seq”, then the value consists of the amino acid letters from Table 3 (capitals for the L-configuration and lower case for the D-configuration), including the letter X for a non-standard amino-acid, and any whitespace character for visual grouping.

14.2.8.7 If value media type is “application/x-na-seq”, then the value consists of the nucleic acid letters from Table 4 (capitals for DNA and lower case for RNA), including the letter X for a non-standard nucleotide, and any whitespace character for visual grouping.

14.2.8.8 If value media type is “application/x-inchi-key”, “application/x-mdl-molfile”, “application/x-smiles”, then there is also another structure characteristic with value media type “application/x-inchi” correctly encoding the structure.

14.2.8.9 If value media type is “application/x-inchi”, then there is another structure characteristic with value media type “application/x-inchi-key” matching the InChI structure.

14.2.8.10 If value media type is “application/x-mdl-molfile”, then its atom list is sorted in the same order as in the other structure characteristic with value media type “application/x-inchi”.

14.2.9 Stereochemistry

```
<moiety>
  <partMoiety><!-- Moiety Details --></partMoiety>
  <subjectOf>
    <characteristic>
      <code code="C18188" displayName="Stereochemistry Type"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>
      <value xsi:type="CV" code="C103211"
             displayName="Square Planar 1"
             codeSystem="2.16.840.1.113883.3.26.1.1"/>
```

The following Table 5 lists the codes for stereochemistry types:

Table 5: Stereochemistry Type

Stereochemistry type	Code
Square Planar 1	C103211
Square Planar 2	C103212
Square Planar 3	C103213
Square Planar 4	C103214
Octahedral 12	C103216
Octahedral 22	C103217
Octahedral 21	C103218
Axial R	C103220
Axial S	C103221

Validation Procedures

14.2.9.1 Code for stereochemistry structure is C18188 and code system is 2.16.840.1.113883.3.26.1.1.

14.2.9.2 There is a coded value (type CV).

14.2.9.3 Value code system is 2.16.840.1.113883.3.26.1.1.

14.2.9.4 Value comes from the Stereochemistry Type Code list.

14.2.9.5 There is no more than one stereochemistry specification per moiety.

14.2.9.6 If Stereochemistry specification characteristic is present, then the moiety must have characteristic "Chemical Structure".

14.2.9.7 If Stereochemistry specification characteristic is present, then the moiety does not have nested moieties.

14.2.10 Optical Activity

When stereochemistry can not be specified specifically in the Chemical Structure, the Optical activity may be specified for the overall substance as follows:

```
<identifiedSubstance>
  <identifiedSubstance> ... </identifiedSubstance>
  <subjectOf>
    <characteristic>
      <code code="C103201" displayName="Optical Activity"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>
      <value xsi:type="CV" code="C103202"
             displayName="plus"
             codeSystem="2.16.840.1.113883.3.26.1.1"/>
```

Validation Procedures

14.2.10.1 Optical activity is under a main substance, not under individual structural units.

14.2.10.2 Code for optical activity is C103201 and code system is 2.16.840.1.113883.3.26.1.1.

14.2.10.3 There is a coded value (type CV).

14.2.10.4 Value code system is 2.16.840.1.113883.3.26.1.1.

14.2.10.5 Value code is *plus* (C103202), *minus* (C103203), or *plus/minus* (C103204).

14.2.11 Structural Repeat Units of Polymers

A special type of a structural unit is polymers which in turn consist of Structural Repeat Units (SRU). SRU is a structural fragment that repeats many times (sequentially or randomly) in a non-template-driven polymer. There are different ways of selecting the SRU from the same “infinite” polymeric chain. For example, the seemingly different SRUs in Figure 2 represent the same polymer.

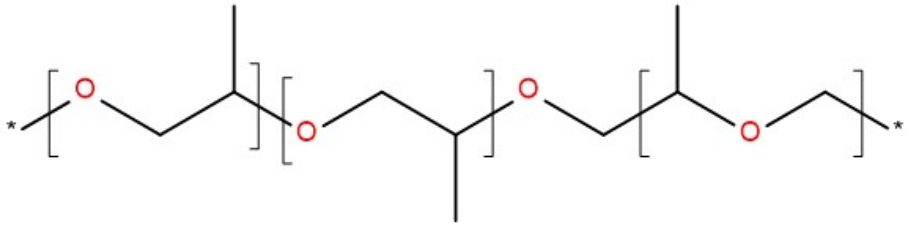


Figure 2: Different SRU candidates for the same polymeric chain.

Homopolymers have only one SRU, but heteropolymers are composed of multiple different SRUs, which is why the SRU(s) of a polymer are nested underneath moiety of type “polymer”.

```
<moiety>
  <code code="C48803" displayName="Polymer"
        codeSystem="2.16.840.1.113883.3.26.1.1"/>
  <quantity ... Quantity Details ... />
  <partMoiety>
    <id extension="A" root="bf170df7-8153-45c1-9a21-6e1aa1b4c97e"/>
    <name>polyethylene glycol 50</name>

    <moiety>
      <code code="C164243" displayName="Structural Repeat Unit"
            codeSystem="2.16.840.1.113883.3.26.1.1"/>

      <quantity>
        <numerator value="50" unit="1"/>
        <denominator value="1" unit="mol" />
      </quantity>

      <partMoiety>
        <id extension="SRU_11"
            root="bf170df7-8153-45c1-9a21-6e1aa1b4c97e"></id>
        <code code="SRU_1"
              codeSystem="bf170df7-8153-45c1-9a21-6e1aa1b4c97e" />
      </partMoiety>
    </moiety>
```

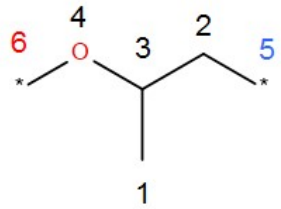


Figure 3: One canonical SRU

Only one *canonical* structure can be selected according to IUPAC rules described in the InChI 1.05 Technical manual on pg. 54-56. This one canonical SRU in our example is shown in Figure 3.

SRUs are defined in an auxiliary substance in the same document after the main substance. Each SRU is defined by its chemical structure and the SRU connection points. The chemical structure has pseudo atoms (*) and no brackets. The SRU structure of Figure 3 is represented by a MOLFILE and uniquely identified by InChI as follows:

```
<subject>
  <identifiedSubstance>
    <id extension="SRU1" root="91432217-1e7b-46f9-ada0-3e09e4d20495" />
    <identifiedSubstance>
      <code code="SRU1"
            codeSystem="91432217-1e7b-46f9-ada0-3e09e4d20495" />
      <asEquivalentSubstance>
        <definingSubstance>
          <code code="6973393c-87f2-75e8-96ad-db924c8d8a44"
                codeSystem="2.16.840.1.113883.3.2705" />
        </definingSubstance>
      </asEquivalentSubstance>
```

```
      <moiety>
        <quantity>
          <numerator value="1" unit="mol" />
          <denominator value="1" unit="mol" />
        </quantity>
        <partMoiety />
        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure" />
            <value xsi:type="ED" mediaType="application/x-mdl-molfile"
><![CDATA[

  -FDASRS-          2D

  6  5  0  0  0  0  0  0  0  0999 V2000
    8.9041  -11.2830    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    9.5281  -10.2023    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    8.9041  -10.5625    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
    8.2802  -10.2023    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   10.1520  -10.5625    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
    7.6563  -10.5625    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0

  1  3  1  0  0  0  0
  2  3  1  0  0  0  0
  2  5  1  0  0  0  0
  3  4  1  0  0  0  0
  4  6  1  0  0  0  0

M  END]]></value>
          </characteristic>
        </subjectOf>

        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure InChI" />
            <value xsi:type="ED" mediaType="application/x-inchi"
              >InChI=1B/C3H6OZz2/c1-3(2-5)4-6/h3H,2H2,1H3</value>
          </characteristic>
        </subjectOf>

        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure InChIKey" />
            <value xsi:type="ED" mediaType="application/x-inchi-key"
              >HWXRSRVZRSEKDJ-UHFFFAOYBA-N</value>
```

SRUs can be *linear*, *branched* or *crosslinked*. The connection points are identified as pseudo-atoms. The connection points are classified into “heads” and “tails”. Head-to-tail connections are always assumed (reading from left to right this means the tail from the left SRU connects to the head of the next right SRU and so on.

The polymer in Figure 2 is comprised of the **linear SRU** in Figure 3. A linear SRU with two pseudo atoms has one head and one tail. The head is the pseudo-atom (*) at 6 adjacent to the oxygen (O) at 4 and the tail is the pseudo-atom (*) at 5 adjacent to the carbon (C) at 2. These SRU connection points are specified as follows:

```
<moiety>
  <code code="C132921" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="Linear SRU Connection Points"/>

  <positionNumber value="6"/>
  <positionNumber value="5"/>

  <partMoiety/>
</moiety>
```

When the SRUs are connected to form the polymer, these pseudo-atoms disappear, leaving the connection in this case from tail carbon (C) at 2 to the head oxygen (O) at 4.

There are **linear SRUs with more than two connection points**, for example, a linear SRU with four connections points has two pairs of head/tail positions, which are connected in parallel to a ladder formation (ladder-polymer).

An example for the SRU of a ladder polymer is shown in Figure 4. Here, the first head/tail pair is the pseudo-atom (*) at 8 connected to the oxygen (O) at 4 and the pseudo-atom (*) at 10 connected to the silicon Si at (6), the second head/tail pair is the pseudo-atom (*) at 9 connected to the oxygen (O) at 5 and the pseudo-atom (*) at 11 connected to the silicon (Si) at 7. That implies SRU-SRU connections in the homopolymer are Si at 6 to O at 4, and Si at 7 to O at 5, strictly parallel, as a ladder.

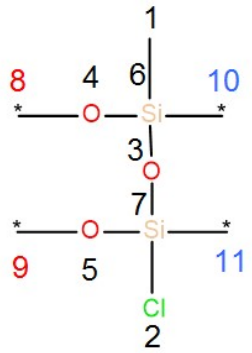


Figure 4: Linear SRU with 2 pairs of head / tail

The SRU connection points for such a ladder polymer are specified in alternating coordinated head-tail pair:

```
<moiety>
  <code code="C132921" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="Linear SRU Connection Points"/>

  <positionNumber value="8"/>  <!-- head -->
  <positionNumber value="10"/> <!-- tail -->

  <positionNumber value="9"/>  <!-- head -->
  <positionNumber value="11"/> <!-- tail -->
</moiety>
```

Thus linear SRU connection points always come as a list of even-numbered elements.

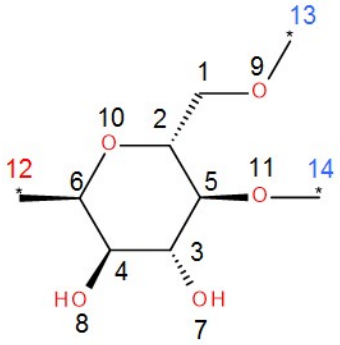


Figure 5: Branched SRU

A branched SRU has one head and more than one tail. For example, in the branched SRU of Figure 5, the head is the pseudo-atom (*) 12 adjacent to ring carbon (C) at 6, the tails are the pseudo-atoms (*) 13 and 14 adjacent to the two oxygens (O) at 9 and 11, respectively. The SRU-SRU connections in a homopolymer are the tail O at 11 to the head C at 6 and the other tail O at 9 to the head C at 6, branching out in a tree-like structure.

The branched SRU connection points are specified as first the one head and then all the tails as in the following example:

```
<moiety>
  <code code="C157466" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="Branched SRU Connection Points"/>
  <positionNumber value="12"/> <!-- head -->
  <positionNumber value="13"/> <!-- tail -->
  <positionNumber value="14"/> <!-- tail -->
</moiety>
```

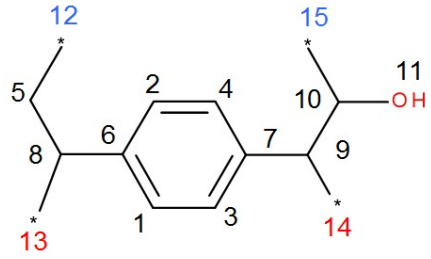


Figure 6: Cross-linked SRU

A cross-linked SRU may have multiple heads and multiple tails. In the example of Figure 6, there are two heads and two tails. The heads are the pseudo-atom (*) 13 and 14 adjacent to carbons (C) at 8 and 9, respectively, and the tails are the pseudo-atoms (*) 12 and 15 adjacent to carbons (C) at 5 and 10, respectively. In the absence of other structural units, the allowed SRU-SRU connections are C at 5 to C at 8, C at 5 to C at 9, C at 10 to C at 8, and C at 10 to C at 9.

The SRU connection points for cross-linked SRUs are specified by first giving all the heads and then all the tails. Because there can be more than two heads, the first position number specifies the number of heads. In this example:

```
<moiety>
  <code code="C164241" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="Cross-linked SRU Connection Points"/>
  <positionNumber value="2"/> <!-- number of heads -->
  <positionNumber value="13"/> <!-- heads -->
  <positionNumber value="14"/> <!-- heads -->
  <positionNumber value="12"/> <!-- tail -->
  <positionNumber value="15"/> <!-- tail -->
</moiety>
```

The heads and tails must be ordered by ascending position number!

In this specification, polymers are always stochastic combinations of the repeated instances of the SRUs. If there are any regular patterns, such as, alternating SRUs, e.g. “A-B-A-B-A-…-B-A-B”, then one can join the two SRUs “-A-” and “-B-” together into one SRU “-A-B-”. Block polymers, such as “A-A-…-A-A-B-B-…-B-B” would be defined as two polymers “A-A-…-A-A” and “B-B-…-B-B” linked together.

Validation Procedures

14.2.11.1 SRU connection points have a moiety code from Table 1 of type “connection points” used with “polymer SRU”.

14.2.11.2 There is one preceding moiety which is the structural unit.

14.2.11.3 The structural unit is a moiety without a code.

14.2.11.4 The structural unit is a moiety that has a chemical structure MOLFILE representation.

14.2.11.5 The structural unit is a moiety that has a chemical structure InChI representation.

14.2.11.6 The MOLFILE has two or more pseudo-atoms represented as asterisks “*”

14.2.11.7 The InChI has the same number of pseudo-atoms, given as symbol “Zz.”

14.2.11.8 There are two or more position numbers.

Linear SRU Connection Points

14.2.11.9 *Linear SRU Connection Points* (C132921) have an even number of position numbers.

14.2.11.10 *Linear SRU Connection Points* (C132921) position numbers are all unique.

14.2.11.11 *Linear SRU Connection Points* (C132921) position numbers all reference a pseudo atom (Zz) in the InChI structure of their structural unit.

14.2.11.12 *Linear SRU Connection Points* (C132921) have position numbers in pairs of head and tail positions, in ascending order of the head atom number, i.e., there is no preceding head position number with a value greater than the current head position number.

Branched SRU Connection Points

14.2.11.13 *Branched SRU Connection Points* (C157466) have one head position number followed by 2 or more tail position numbers.

14.2.11.14 *Branched SRU Connection Points* (C157466) position numbers are all unique.

14.2.11.15 *Branched SRU Connection Points* (C157466) position numbers all reference a pseudo atom (Zz) in the InChI structure of their structural unit.

14.2.11.16 *Branched SRU Connection Points* (C157466) have their tail positions ordered by ascending value, i.e., there is no preceding tail position number with a value greater than the current tail position number.

Cross-Linked SRU Connection Points

14.2.11.17 *Cross-Linked SRU Connection Points* (C164241) have the first position number indicating the number 2 or more of head positions.

14.2.11.18 The position numbers following the first position number of *Cross-Linked SRU Connection Points* (C164241) define the head positions.

14.2.11.19 The number of head positions of *Cross-Linked SRU Connection Points* (C164241) is equal to the value of the first position number.

14.2.11.20 *Cross-Linked SRU Connection Points* (C164241) have all the tail positions following the head positions.

14.2.11.21 *Cross-Linked SRU Connection Points* (C164241) have at least two tail positions.

14.2.11.22 *Cross-Linked SRU Connection Points* (C164241) head and tail position numbers are all unique.

14.2.11.23 *Cross-Linked SRU Connection Points* (C164241) head and tail position numbers all reference a pseudo atom (Zz) in the InChI structure of their structural unit.

14.2.11.24 *Cross-Linked SRU Connection Points* (C164241) head positions are ordered by ascending value.

14.2.11.25 *Cross-Linked SRU Connection Points* (C164241) tail positions are ordered by ascending value.

14.2.12 Irregular Substituents and R-Groups

All post-translational modifications and other variances from the closest amino acid sequence are specified by means of amino acid substitutions. All variances from the natural nucleotide sequence are specified by means of nucleotide substitutions. All end groups and some other modifications of non-template driven polymers are specified by means of SRU substitutions. R-Groups are used for Markush structures and are explained in Section 14.2.17.

In an amino acid substitution one regular amino acid is replaced by an irregular amino acid, or any molecule that fits into the amino acid chain. Such a molecule must be defined by a chemical structure and Amino Acid Connection Points. This means, usually, one amino group must be marked to substitute the amino group of the original amino acid, and one carboxyl group must be defined to substitute the carboxyl group of the original amino acid. However other configurations are possible.

Terminal substitutions may only need one Amino Acid Connection Point, a carboxyl group for N-terminal and an amino group for C-terminal substitutions.

There may be more than two Amino Acid Connection Points to provide for cross-links between chains or sections of one chain. For example, in the disulfide bridge, two cysteine (C) amino acids are replaced by one shared *cystine* integrated in both chains (Figure 7).

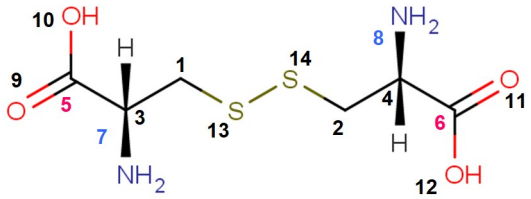
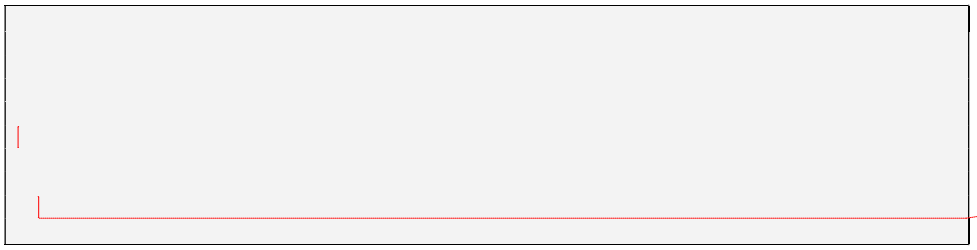


Figure 7: Disulfide bridge, a cystine, irregular amino acid with two pairs of amino acid connection points..

The cystine, for example, would be defined as an auxiliary locally defined substance with Amino Acid Connection Points as follows:



Comment [GS1]: [RESERVED]

```
      <moiety>
        <quantity>
          <numerator value="1" unit="mol" />
          <denominator value="1" unit="mol" />
        </quantity>
        <partMoiety>
          <code code="48TCX9A1VT" codeSystem="2.16.840.1.113883.4.9"/>
        </partMoiety>
        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure"/>
            <value xsi:type="ED"
                   mediaType="application/x-mdl-molfile" >...</value>
          </characteristic>
        </subjectOf>

        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure InChI"/>
            <value xsi:type="ED" mediaType="application/x-inchi">
              InChI=1S/C6H12N2O4S2/c7-3(5(9)10)1-13-14-2-4(8)6(11)12/h3-
4H,1-2,7-8H2,(H,9,10)(H,11,12)</value>
          </characteristic>
        </subjectOf>

        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure InChIKey"/>
            <value xsi:type="ED" mediaType="application/x-inchi-key">
              LEVWYRKDKASIDU-UHFFFAOYSA-N</value>
          </characteristic>
        </subjectOf>
```

After InChI normalization the atoms are uniquely numbered as indicated in Figure 7, and based on these canonical InChI atom numbers, the Amino Acid Connection Points are defined. The carboxyl group with carbon (C) 5 and the amino group with nitrogen (N) at 7 constitute one pair of amino acid connectin points, and the carboxyl group with carbon (C) at 6 and the amino group with nitrogen (N) at 8 constitute another pair of Amino Acid Connectin Points.

```
      <moiety>
        <code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1"
              displayName="Amino Acid Connection Points"/>
        <positionNumber value="7"/>
        <positionNumber value="5"/>
        <partMoiety/>
      </moiety>

      <moiety>
        <code code="C118427" codeSystem="2.16.840.1.113883.3.26.1.1"
              displayName="Amino Acid Connection Points"/>
        <positionNumber value="8"/>
        <positionNumber value="6"/>
        <partMoiety/>
      </moiety>
```

This completes the definition of a cystin disulfide bridge, which now can be re-used many times to describe many disulfide bridges between and within various protein subunits.

The two position number elements referencing InChI canonicalized atom numbers have a designated meaning depending on the connection point moiety code, as summarized in the following Table 6.

Table 6: Connection Points Numbering

Moiety Code	Name	1st Position	Atom	2nd Position	Atom
C118427	Amino acid connection point	Amino group	N	Carboxyl group	C
C157467	Nucleotide connection point	Phosphate or thio-phosphate group	P	3' Hydroxyl group of the sugar or whatever atom may substitute it in the substituent.	(O)
C132921	Linear SRU Connection Points	Head	*	Tail	*
C177645	R-Group Connection Point	Attachment point	*	*(none)*	*(none)*

All Connection Points must be sorted by ascending first position number, and, if the first position number is not applicable, the second position number.

The chemical structure of irregular substituents of polymer SRUs and R-Groups (see Section 14.2.18) are given as fragments, i.e., in the MOLFILE drawings, there would be asterisks, "star-atoms", or pseudo atoms, and in the InChI expressions those appear as "Zz" atoms. Conversely, the chemical structures of irregular substituents in sequences of amino acids or nucleotides are given as monomer molecules, not as fragments.

Because polymers are specified using SRUs which are always given as fragments, when the end groups are known they are specified as *SRU substitutions* with irregular substituents on position 1 (replacing the beginning SRU) and -1 (replacing the terminal SRU).

Validation Procedures

14.2.12.1 Substituent connection points moiety codes are from Table 6: Connection Points Numbering.

14.2.12.2 There is one preceding moiety element with a moiety code of type "structural unit" (see Table 1) or no code

14.2.12.3 There are one or more. position numbers.

14.2.12.4 At least one of the position number elements has a positive integer number value.

14.2.12.5 If the structure is not intended to have one of the two functional groups (e.g., an amino group or a carboxyl group), then the respective position number has no value but a nullFlavor attribute set to “NA” (i.e., not applicable).

14.2.12.6 The first position number is greater than the first position number of the preceding connection points moiety, if any.

14.2.12.7 If the first position number is not applicable, then the first position number of the preceding connection point moiety, if any, is also not applicable.

14.2.12.8 If the first position number is not applicable, then the second position number is greater than the preceding one’s.

14.2.12.9 A connection point moiety is preceded by the structural unit with an InChI structure representation.

Amino-Acid Connection Points

14.2.12.10 There are two position numbers.

14.2.12.11 If the moiety code is *Amino Acid Connection Points* (C118427), then the first position number references the nitrogen atom (N) of the amino group.

14.2.12.12 If the moiety code is *Amino Acid Connection Points* (C118427), then the second position number references the carbon atom (C) of a carboxyl group.

Nucleotide Connection Points

14.2.12.13 There are two position numbers.

14.2.12.14 If the moiety code is *Nucleotide Connection Points* (C157467), then the first position number references the phosphorus atom (P) of a phosphate or thio-phosphate group.

14.2.12.15

Linear SRU Connection Points

14.2.12.16 *Linear SRU Connection Points* (C132921) have an even number of position numbers.

14.2.12.17 *Linear SRU Connection Points* (C132921) position numbers are all unique.

14.2.12.18 *Linear SRU Connection Points* (C132921) position numbers all reference a pseudo atom (Zz) in the InChI structure of their structural unit.

14.2.12.19 *Linear SRU Connection Points* (C132921) have position numbers in pairs of head and tail positions, in ascending oder of the head atom number, i.e., there is no preceding head position number with a value greater than the current head position number.

R-Group Connection Points

14.2.12.20 *R-Group Connection Point* (C177645) have one position number.

14.2.12.21 *R-Group Connection Point* (C177645) position numbers reference a pseudo atom (Zz) in the InChI structure of their structural unit.

14.2.13 Structural Modifications

Any molecule can have structural modifications if those structural modifications could not simply be drawn as a different molecule. That is, simple chemicals that are defined as a single structural units do not have just determined modifications. But if there are any repeating groups or sequences, then the only way to describe irregularities are as such Structural Modifications. Structural modifications are represented as moiety elements that follow the basic structural units of the substance.

```
<identifiedSubstance>
  ...
  <identifiedSubstance>
    ...
    <moiety><!-- Structural Unit Moiety --></moiety>
    ...
    <moiety><!-- Structural Modification Moiety -->
      <code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="STRUCTURAL MODIFICATION" />
      ...
    </moiety>
    ...
```

A structural modification moiety normally has one or more bond elements which indicate where the structural modification occurs in the base structure.

```
<moiety>
  <code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="STRUCTURAL MODIFICATION" />
  <partMoiety>
    ...
    <bond>
      <code code="Bond Element Type Code"
            codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Bond Elememt Type Name"/>
      <positionNumber value="..."/>
      ...
      <distalMoiety>
        <id extension="Id of a Structural Unit" root="Document Id"/>
      </distalMoiety>
    </bond>
```

The bond elements type indicates what type of modification occurs at the site.

Table 7: Bond Element Types

Code	Display Name
C118426	Amino Acid Substitution Site
C157224	Nucleotide Substitution Site
C132923	Structural Repeat Unit Substitution Site
C14050	Structural Attachment Site
C177524	R-Group
C165845	Monomer Deletion Site

Validation Procedures

14.2.13.1 Code for *Structural Modification* is C118425.

14.2.13.2 Code system is 2.16.840.1.113883.3.26.1.1.

14.2.13.3 Display name matches the code.

14.2.13.4 There is a preceding structural unit moiety elements with a moiety code of type "structural unit" (see Table 1) or no code.

14.2.13.5 There is a part moiety element, representing the modification.

14.2.13.6 There is an id with root being the document id.

14.2.13.7 Id is unique in the entire document.

14.2.13.8 A structural modification has one or more bond elements of type listed in Table 7.

14.2.14 Monomer Substitutions

Substituents identified locally in the document (see Section 14.2.12) can then be referenced by actual monomer substitutions.

Amino acid substitutions are specified as moieties of code "structural modification" (C118425) that contain one or more bond elements of type "Amino Acid Substitution Site" (C118426). The moiety substance (partMoiety) is referenced by the code element (e.g., "cys-cys" in the example below), which must cite an irregular amino acid substance for substitution (see Section 14.2.12) locally defined in the document, with the appropriately matching amino acid connection point moieties specified. All amino acid substitutions must also have their own unique id (e.g. here "BR1" in the example below).

```
<moiety>
  <code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="STRUCTURAL MODIFICATION" />
  <partMoiety>
    <id extension="BR1" root="Document Id"/>
    <code code="cys-cys" codeSystem="Document Id"
          displayName="Cysteine disulfide"/>

    <bond>
      <code code="C118426" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="AMINO ACID SUBSTITUTION SITE"/>
      <positionNumber value="1"/>
      <positionNumber value="53"/>
      <distalMoiety>
        <id extension="SU1" root="Document Id" />
      </distalMoiety>
    </bond>

    <bond>
      <code code="C118426" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="AMINO ACID SUBSTITUTION SITE"/>
      <positionNumber value="2" />
      <positionNumber value="62" />
      <distalMoiety>
        <id extension="SU1" root="Document Id"/>
      </distalMoiety>
    </bond>
```

The bond elements connect the irregular amino acid to the correct position on the specified protein subunit (e.g., “SU1” in above example). Each substitution is specified with one bond. In the case of bridges between two chains (or two regions of the same chain), there are two bond elements, one for each side of the bridge.

The same logic applies to modifications of nucleic acids and other polymers. Such substitutions are specified as moieties of code “Structural modification” (C118425) with bond elements of type “Nucleotide Substitution Site (C157224) and “Structural Repeat Unit Substitution Site” (C132923) correspondingly.

Validation Procedures

14.2.14.1 There are one or more preceding moiety elements with a moiety code of type “structural unit” (see Table 1) or no code.

14.2.14.2 There is no following structural unit moiety element.

14.2.14.3 There is a partMoiety element

14.2.14.4 The partMoiety element has a code

14.2.14.5 The part moiety code references the code of a *Substituent* (see Section 14.2.12) defined locally in this document.

14.2.14.6 The Substituent has one or more *Connection Point Moieties*.

14.2.14.7 The partMoiety element has one or more bond elements.

14.2.14.8 Bond element code is from Table 7: Bond Element Types except *Monomer Deletion Site* (C165845) or *Structural Attachment Site* (C14050).

14.2.14.9 Code system is 2.16.840.1.113883.3.26.1.1.

14.2.14.10 Display name matches the code.

14.2.14.11 There are two position number elements.

14.2.14.12 There is a distal moiety element.

14.2.14.13 Distal moiety element has an id.

14.2.14.14 Distal moiety id root is the same as the document id root.

14.2.14.15 Distal moiety extension refers to a preceding structural unit moiety with the same id.

14.2.14.16 The distal moiety so referenced by the id has a chemical structure characteristic, except if the bond code is *Structural Repeat Unit Substitution Site* (C132923).

Amino Acid Substitution Site (C118426)

14.2.14.17 For an *Amino Acid Substitution Site* (C118426), the *Substituent* is an amino acid, i.e., it has one or more *Amino Acid Connection Points* (C118427).

14.2.14.18 For *Amino Acid Substitution Site* (C118426), the first position number selects an *Amino Acid Connection Points* (C118427) moiety of the *Substituent* (see Section 14.2.12).

14.2.14.19 For an *Amino Acid Substitution Site* (C118426), the distal moiety refers to a *Protein Subunit* (C118424).

14.2.14.20 For an *Amino Acid Substitution Site* (C118426), the distal moiety (*Protein Subunit,* C118424) has a structure representation of type amino acid sequence (mediaType is “application/x-aa-seq”)

14.2.14.21 For an *Amino Acid Substitution Site* (C118426), the distal moiety (*Protein Subunit,* C118424) has amino acid sequence length of at least 1.

14.2.14.22 For an *Amino Acid Substitution Site* (C118426), the second position number is greater or equal to 1.

14.2.14.23 For an *Amino Acid Substitution Site* (C118426), the second position number is a position on the amino acid sequence of the protein subunit referenced by the distal moiety, and therefore is less or equal to the length of that sequence.

14.2.14.24 If the referenced *Amino Acid Connection Points* moiety of the *Substituent* (see Section 14.2.12) has no amino group (i.e., its first position number is N/A), then the second position number of this *Amino Acid Substitution Site* is equal to 1.

14.2.14.25 If the referenced *Amino Acid Connection Points* moiety of the *Substituent* (see Section 14.2.12) has no carboxyl group (i.e., its second position number is N/A), then the second position number of this *Amino Acid Substitution Site* is equal to the length of the amino acid sequence.

Nucleotide Substitution Site (C157224)

14.2.14.26 For a *Nucleotide Substitution Site* (C157224), the *Substituent* is a nucleotide, i.e., it has one or more *Nucleotide Connection Points* (C157467).

14.2.14.27 For a *Nucleotide Substitution Site* (C157224), the first position number selects a *Nucleotide Connection Points* (C157467) moiety of the *Substituent* (see Section 14.2.12).

14.2.14.28 For a *Nucleotide Substitution Site* (C157224), the distal moiety refers to a *Polynucleotide* (C164242).

14.2.14.29 For a *Nucleotide Substitution Site* (C157224), the distal moiety (*Polynucleotide*, C164242) has a structure representation of type nucleic acid sequence (mediaType is “application/x-na-seq”)

14.2.14.30 For a *Nucleotide Substitution Site* (C157224), the distal moiety (*Polynucleotide*, C164242) has nucleotide sequence length of at least 1.

14.2.14.31 For a *Nucleotide Substitution Site* (C157224), the second position number is greater or equal to 1.

14.2.14.32 For a *Nucleotide Substitution Site* (C157224), the second position number is a position on the nucleotide sequence of the polynucleotide referenced by the distal moiety, and therefore is less or equal to the length of that sequence.

14.2.14.33 If the referenced *Nucleotide Connection Points* moiety of the *Substituent* (see Section 14.2.12) has no phosphate group (i.e., its first position number is N/A), then the second position number of this *Nucleotide Substitution Site* is equal to 1.

14.2.14.34 If the referenced *Nucleotide Connection Points* moiety of the *Substituent* (see Section 14.2.12) has no 3' sugar hydroxyl group (i.e., its second position number is N/A), then the second position number of this *Nucleotide Substitution Site* is equal to the length of the nucleic acid sequence.

Structural Repeat Unit Substitution Site (C132923)

14.2.14.35 For a *Structural Repeat Unit Substitution Site* (C132923), the *Substituent* is an irregular SRU, i.e., it has one or more *Linear SRU Connection Points* (see Table 6).

14.2.14.36 For a *Structural Repeat Unit Substitution Site* (C132923), the first position number selects a *Polymer SRU Connection Points* moiety of the *Substituent* (see Section 14.2.12).

14.2.14.37 For a *Structural Repeat Unit Substitution Site* (C132923), the distal moiety refers to a *Polymer* (C48803).

14.2.14.38 For a *Structural Repeat Unit Substitution Site* (C132923), the second position number is a position on the polymer referenced by the distal moiety, and can have the values 1 for first SRU, -1 for last or the nullFlavor *unknown* (UNK) for unknown/undetermined SRU positions.

14.2.15 Monomer Deletion

Monomer deletion is a complete removal of the entire monomer (an amino acid or a nucleotide) from a biopolymer. Monomer deletions are specified as moieties of code "Structural Modification" (C118425) which contain one bond element of type "Monomer Deletion Site" (C165845).

```
<moiety>
  <code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="Structural Modification" />
  <partMoiety>
    <bond>
      <code code="C165845" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Monomer Deletion Site"/>
      <positionNumber value="100"/>
      <distalMoiety>
        <id extension="SU1" root="Document Id" />
      </distalMoiety>
    </bond>
```

The bond element indicates location of the deletion in a polypeptide or in a polynucleotide subunit.

Validation Procedures

14.2.15.1 There are one or more preceding moiety elements with a moiety code of type "structural unit" (see Table 1) or no code.

14.2.15.2

14.2.15.3 There is no following structural unit moiety element

14.2.15.4 There is a partMoiety element

14.2.15.5 There is no code in partMoiety.

14.2.15.6 Bond code is *Monomer Deletion Site* (C165845, code system 2.16.840.1.113883.3.26.1.1).

14.2.15.7 There is one position number element.

14.2.15.8 There is a distal moiety element.

14.2.15.9 Distal moiety element has an id.

14.2.15.10 Distal moiety id root is the same as the document id root.

14.2.15.11 Distal moiety extension refers to a preceding structural unit moiety with the same id.

14.2.15.12 The preceding structural unit moiety so referred to by the distal moiety id is a *Protein Subunit* (C118424) or a *Polynucleotide* (C164242).

14.2.15.13 The preceding structural unit moiety so referred to by the distal moiety id has a chemical structure characteristic.

14.2.15.14 The chemical structure characteristic is an amino acid sequence or nucleic acid sequence (mediaType "application/x-aa-seq" or "application/x-na-seq", respectively).

14.2.15.15 The distal moiety refers to a sequence of length of at least 2.

14.2.15.16 The position number is greater or equal to 1.

14.2.15.17 The position number is a valid ordinal position in the referenced sequence and therefore less or equal to the length of that sequence.

14.2.16 Structural Attachment Modification

Structural attachment modifications on proteins are used if the attachment is not fully structurally defined so that one cannot use an amino acid substitution. Structural attachment modifications have one bond element connecting to the protein subunit and specifying the position to which the attachment is connected. The specific structure and the point of attachment are not specified explicitly but may sometimes be determined from the attachment type and the amino acid at the indicated position.

Structural attachment modifications are used to specify glycosylations. The exact structure of the glycan is not known, only specified by the following glycan types:

Table 8: Glycan Types

NCIt Code	Name	NCIt Code	Name
C118429	Avian Type Glycan	C128564	Plant Type Glycan
C118430	Bacterial Type Glycan	C158047	Insect Type Glycan
C118432	Fungal Type Glycan	C158048	Reptilian Type Glycan
C118428	Human Type Glycan	C158049	Mammalian Afucosylated Glycan
C118431	Mammalian Type Glycan	C158050	Molluscan Type Glycan

The following is an example for a glycosylation structural attachment modification:

```
<moiety>
  <code code="C118425" codeSystem="2.16.840.1.113883.3.26.1.1"
        displayName="STRUCTURAL MODIFICATION" />
  <partMoiety>
    <id extension="GLY1" root="Document Id"/>
    <code code="C118430" codeSystem="2.16.840.1.113883.3.26.1.1"
          displayName="Bacterial Type Glycan"/>

    <bond>
      <code code="C14050" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="STRUCTURAL ATTACHMENT SITE"/>
      <positionNumber value="153"/>
      <distalMoiety>
        <id extension="SU1" root="Document Id" />
      </distalMoiety>
    </bond>
```

The bond element connects the attachment modification to the correct position on the specified protein subunit (e.g., “SU1” in the example above).

Validation Procedures

14.2.16.1 There is a preceding protein subunit moiety element.

14.2.16.2 There is no following protein subunit moiety element.

14.2.16.3 There is a code with code system 2.16.840.1.113883.3.26.1.1.

14.2.16.4 Code is from the glycan type list.

14.2.16.5 Display name matches the code.

14.2.16.6 There is one amino acid substitution site bond element.

14.2.16.7 Code is C14050.

14.2.16.8 Code system is 2.16.840.1.113883.3.26.1.1.

14.2.16.9 Display name matches the code.

14.2.16.10 There is one position number element.

14.2.16.11 There is a distal moiety element with an id locally defined in the document, i.e., the id root is the document id.

14.2.16.12 Distal moiety id references a protein subunit by its identifier locally defined in a preceding moiety.

14.2.16.13 Position number references an amino acid position on the referenced protein subunit and therefore is greater or equal to 1.

14.2.16.14 Position number references an amino acid position on the referenced protein subunit and therefore is less or equal to the number of amino acids in the protein subunit.

14.2.17 Markush Structures / R-Group Modification

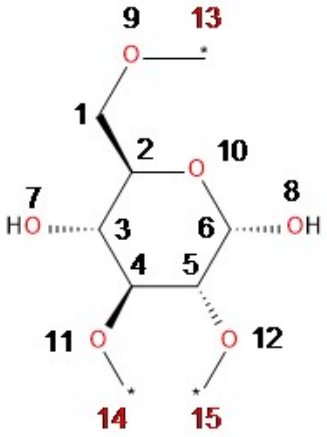


Figure 8: Sugar with Markush Attachments.

Markush structures describe a compound class by generic notation:

- Substitution variation (R-groups, atom and bond lists)
- Frequency variation (link nodes and repeating units)
- Position variation (variable point of attachment)
- Homology variation (e.g. alkyl, aryl)
- Conditions for generic features: occurrence lists, dependency, etc.

Markush groups are used when we cannot simply enumerate the variant structures in one flat list. Instead Markush groups can “generate” a large number of

variant structures. They are only used here when we *both* the position and the attached structures at each position undetermined, i.e., if there is more than one possible site at which more than one possible structure may be attached.

For example, the sugar in Figure 8 has 3 possible attachment sites, and that is represented as follows:

```
<subject>
  <identifiedSubstance>
    <identifiedSubstance>
      <code code="SU" codeSystem="Document Id"/>
      ...
      <moiety>
        ...
        <partMoiety>
          <id extension="SU" codeSystem="Document Id"/>
        </partMoiety>
        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure" />
            <value xsi:type="ED" mediaType="application/x-mdl-molfile"
><![CDATA[

  ACCLDraw02152117082D

     RDKit          2D

17 17  0  0  1  0  0  0  0  0999 V2000

   12.0757   -4.9639    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.6662   -5.9868    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.0756   -7.0096    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.6662   -8.0325    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.8473   -8.0325    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.4379   -7.0096    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   10.8945   -7.0096    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   15.6190   -7.0096    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   12.6662   -3.9411    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   13.8473   -5.9868    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   12.0756   -9.0554    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   14.4379   -9.0554    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   13.8473   -3.9411    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
   12.6662  -10.0782    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
   13.8473  -10.0782    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
   10.5645   -7.0096    0.9334 H   0  0  0  0  0  0  0  0  0  0  0  0
   15.9490   -7.0096   -0.9334 H   0  0  0  0  0  0  0  0  0  0  0  0
```

```
  3  7  1  6
  3  2  1  0
  2 10  1  0
 10  6  1  0
  6  8  1  6
  6  5  1  0
  5  4  1  0
  4  3  1  0
  4 11  1  1
 11 14  1  0
  5 12  1  6
 12 15  1  0
  2  1  1  1
  1  9  1  0
  9 13  1  0
  7 16  1  0
  8 17  1  0
M  END]]></value>
          </characteristic>
        </subjectOf>

        <subjectOf>
          <characteristic>
            <code code="C103240" codeSystem="2.16.840.1.113883.3.26.1.1"
                  displayName="Chemical Structure InChI" />
            <value xsi:type="ED" mediaType="application/x-inchi"
              >InChI=1B/C6H9O6Zz3/c7-3-2(1-9-13)10-6(8)5(12-15)4(3)11-
14/h2-8H,1H2/t2-,3-,4+,5-,6+/m1/s1</value>
          </characteristic>
        </subjectOf>
        ...
```

Now the actual Markush attachment modifications are represented as structural modification elements as has already been defined in Section 14.2.13.

```
      <moiety>
        <code code="C118425" displayName="STRUCTURAL MODIFICATION"
              codeSystem="2.16.840.1.113883.3.26.1.1"/>
        <quantity>
          <numerator value="0.5" unit="mol" />
          <denominator value="1" unit="mol" />
        </quantity>
        <partMoiety>
          <id extension="M1" root="Document ID" />
          <code code="R1" codeSystem="Document ID" />
          <bond>
            <code code="C177524" displayName="R-group"
                  codeSystem="2.16.840.1.113883.3.26.1.1"/>
            <positionNumber value="1"  R-Group Connection Point -->

            <positionNumber value="13" possible atom number in SU />
            <positionNumber value="14" possible atom number in SU />
            <positionNumber value="15" possible atom number in SU />

            <distalMoiety>
              <id extension="SU" root="Document ID"/>
            </distalMoiety>
```

The bond element of type *R-group* (C177524) links the R-group to the structural unit moiety. The R-group must be defined as an auxiliary substance somewhere in the document, and is referred to by the code (here “R1” with the codeSystem always

being the document id, which is a UUID). The structural modification moiety's partMoiety element also has an id which is unique to this attachment modification; so that if there was another structure elsewhere in the document that has the same attachment group, the code "R1" would be the same, but the id would be different (e.g. "M2"). See Section 14.2.7 about the code and id. The structural unit moiety is the distal moiety of the modification bond and is referred to by its id "SU". That id "SU" was defined earlier in the structural unit moiety's partMoiety element (one can think of those ids as "labels").

The sequence of position numbers are interpreted as follows.

First position number refers to a R-Group Connection Point(s) moiety of the R-Group definition.

The next position numbers refer to the canonical InChI atom numbers of the structural unit referenced by the distalMoiety id, where each atom pointed to is a candidate attachment site. There are always more than one candidate attachment sites, or else we wouldn't use a Markush expression.

If there is one structural modification with bond type R-Group, then there must always be another one referring to a different R-Group:

```
      <moiety>
        <code code="C118425" displayName="STRUCTURAL MODIFICATION"
              codeSystem="2.16.840.1.113883.3.26.1.1"/>
        ...
        <partMoiety>
          <id extension="M2" root="Document ID" />
          <code code="R2" codeSystem="Document ID" />
          <bond>
            <code code="C177524" displayName="R-group"
                  codeSystem="2.16.840.1.113883.3.26.1.1"/>
            <positionNumber value="1" atom position in R-group />
            <positionNumber value="13" possible atom number in SU />
            <positionNumber value="14" possible atom number in SU />
            <positionNumber value="15" possible atom number in SU />
            <distalMoiety>
              <id extension="SU" root="Document ID"/>
            </distalMoiety>
```

In this example, the position numbers are all the same, but this need not be so. It is possible to say that some R-Groups attach to some sites, while other R-Groups attach to some of the same and some other sites. But for each possible attachment site, there would always be more than one candidate R-Group to attach, or else the indeterminism goeas away and the molecule at that site can be drawn explicitly.

R-Groups used for Markush structures are auxiliary chemical substances, defined just like any other chemical substance, i.e., with a code and a structure representation. The following example defines the hydroxypropyl R-group of Figure 9. The first part is no different than any other substance specification, but then followed by a R-Group Connection Point moiety that points at the star-atom in the structure.

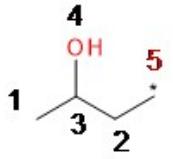


Figure 9: Hydroxypropyl R-Group

```
<subject>
  <identifiedSubstance>
    <identifiedSubstance>
      <code code="R1" codeSystem="bf170df7-8153-45c1-9a21-6e1aa1b4c97e"/>
      <moiety>
        ...
        <partMoiety/>

        <subjectOf>
          <characteristic>
            <code code="C103240" displayName="Chemical Structure"
                  codeSystem="2.16.840.1.113883.3.26.1.1"/>
            <value xsi:type="ED" mediaType="application/x-mdl-molfile"
><![CDATA[
     RDKit          2D

  6  5  0  0  1  0  0  0  0  0999 V2000

   13.2171   -7.2289    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.6377   -7.2248    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.9274   -6.8136    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.9274   -5.9912    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   15.3481   -6.8093    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
   13.9274   -5.6612    0.9334 H   0  0  0  0  0  0  0  0  0  0  0  0

  2  5  1  0
  3  2  1  0
  3  1  1  0
  3  4  1  0
  4  6  1  0

M  END]]></value>
          </characteristic>
        </subjectOf>
        <subjectOf>
          <characteristic>
            <code code="C103240" displayName="Chemical Structure"
                  codeSystem="2.16.840.1.113883.3.26.1.1" />
            <value xsi:type="ED" mediaType="application/x-
inchi">InChI=1B/C3H7OZz/c1-3(4)2-5/h3-4H,2H2,1H3</value>
          </characteristic>
        </subjectOf>
      </moiety>

      <moiety>
        <code code="C177645" displayName="R-Group Connection Point"
              codeSystem="2.16.840.1.113883.3.26.1.1"/>
        <positionNumber value="5"/>
        <partMoiety/>
      </moiety>
```

The position number element in the R-Group Connection Points moiety element, is referring to the InChI canonical atom number where the attachment will connect.

Even as simple an R-Group as hydrogen are explicitly specified, as follows:

```
<subject>
  <identifiedSubstance>
    <identifiedSubstance>
      <code code="R2" codeSystem="Document Id" />
      <moiety>
        ...
        <partMoiety/>
        <subjectOf>
          <characteristic>
            <code code="C103240" displayName="Chemical Structure"
                  codeSystem="2.16.840.1.113883.3.26.1.1"/>
            <value xsi:type="ED" mediaType="application/x-mdl-
molfile"><![CDATA[
     RDKit          2D

  2  1  0  0  1  0  0  0  0  0999 V2000
   14.6300   -7.0625    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
   13.8075   -7.0625    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0
V    1 *
M  END]]></value>
          </characteristic>
        </subjectOf>
        <subjectOf>
          <characteristic>
            <code code="C103240" displayName="Chemical Structure"
                  codeSystem="2.16.840.1.113883.3.26.1.1"/>
            <value xsi:type="ED" mediaType="application/x-
inchi">InChI=1B/HZz/h1H</value>
```

Validation Procedures

14.2.17.1 Markush / R-Group modification are structural modification moieties with one or more R-Group bonds.

14.2.17.2 There is a preceding structural unit moiety element.

14.2.17.3 There is no following structural unit moiety element

14.2.17.4 The structural unit is a moiety without a code.

14.2.17.5 There is a partMoiety element

14.2.17.6 The partMoiety element has an id

14.2.17.7 Id extension is unique in the entire document.

14.2.17.8 Id root is the document id.

14.2.17.9 The partMoiety element has a code

14.2.17.10 The part moiety code references the code of a *R-Group* (see Section 14.2.12) defined locally in this document.

14.2.17.11 The R-Group has one or more *Connection Point Moieties*.

14.2.17.12 The partMoiety element has one or more bond elements.

14.2.17.13 Bond element code is *R-Group* (C177524).

14.2.17.14 Code system is 2.16.840.1.113883.3.26.1.1.

14.2.17.15 Display name matches the code.

14.2.17.16 There are three or more position numbers on each bond element.

14.2.17.17 There is a distal moiety element.

14.2.17.18 Distal moiety element has an id.

14.2.17.19 Distal moiety id root is the same as the document id root.

14.2.17.20 Distal moiety extension refers to a preceding structural unit moiety with the same id.

14.2.17.21 The distal moiety so referenced by the id has a chemical structure characteristic.

14.2.17.22 The structural unit is a moiety that has a chemical structure MOLFILE representation.

14.2.17.23 The MOLFILE has at least as many star-atoms as the count of positionNumber elements minus one.

14.2.17.24 The structural unit is a moiety that has a chemical structure InChI representation.

14.2.17.25 The InChI has the same number of pseudo-atoms as the MOLFILE, given as symbol “Zz.”

14.2.17.26 The first position number references an R-Group Connection Points moiety of the R-Group

14.2.17.27 The other position numbers all reference a pseudo atom (Zz) in the InChI structure of their structural unit.

14.2.18 Authority Citation for Organisms

If the substance is an organism, e.g., a botanical, microorganism, animal, or virus, such organism is defined by its authority citation.

```
<identifiedSubstance>
  <id extension="3R2AD49R6T" root="2.16.840.1.113883.4.9" />
  <identifiedSubstance>
    <code code="3R2AD49R6T" codeSystem="2.16.840.1.113883.4.9" />
    ...
  </identifiedSubstance>
  <subjectOf>
    <document>
      <bibliographicDesignationText>Drimia indica (Roxb.)
Jessop</bibliographicDesignationText>
    </document>
  </subjectOf>
</identifiedSubstance>
```

Validation Procedures

14.2.18.1 There is one bibliographic designation text.

14.2.18.2 Bibliographic designation text is a simple string

14.2.18.3 There are no other elements besides bibliographic designation text

14.2.19 Processes

If the substance cannot be fully defined by its chemical structure or by an authority citation, one can include into the definition the process by which it is produced. All processes begin with a source material, and then list a number of chemical agent or physical modification steps.

```
<identifiedSubstance>
  <identifiedSubstance .../>
  <productOf>
    <quantity value="1" unit="1"/>
    <derivationProcess>
      <code code="C25572" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Modification"/>
      <interactor typeCode="CSM" ... Source Material .../>
```

```
      <component>
        <sequenceNumber value="1"/>
        <derivationProcess>
          <interactor typeCode="CSM" .../> <!-- optional -->
          <controlVariable .../>           <!-- optional -->
        </derivationProcess>
      </component>
      ...
    </derivationProcess>
  </productOf>
</identifiedSubstance>
```

Validation Procedures

14.2.19.1 There may be one productOf element

14.2.19.2 The productOf element may have one quantity element, i.e., the reference amount relative to which all interactor quantities are stated.

14.2.19.3 Quantity has a numeric value.

14.2.19.4 Quantity has a unit.

14.2.19.5 Quantity unit may be mol, 1, g or L.

14.2.19.6 The productOf element has one derivation process element

14.2.19.7 Derivation process code is *Modification* (C25572) of code system 2.16.840.1.113883.3.26.1.1.

14.2.19.8 There is one interactor with type CSM (the source material).

14.2.20 Process Interactor

There are two uses for process interactors. First, the entire process with multiple steps has one interactor as its source material. Then individual process steps may involve other interactors for agents, reagents, catalysts, etc.

```
<derivationProcess>
   ...
   <interactor typeCode="CSM">
     <quantity value="1" unit="1"></quantity>
     <identifiedSubstance>
       <id extension="9W8ZP67K1Z" root="2.16.840.1.113883.4.9"></id>
       <identifiedSubstance>
         <code code="9W8ZP67K1Z" codeSystem="2.16.840.1.113883.4.9"
               displayName="PERSEA AMERICANA WHOLE" />
         <asEquivalentSubstance>
           <definingSubstance>
             <code code="b7b697a0-fe1b-d772-470d-5fd4a1bba445"
                   codeSystem="2.16.840.1.113883.3.2705" />
           </definingSubstance>
         </asEquivalentSubstance>
       </identifiedSubstance>
     </identifiedSubstance>
   </interactor>
```

14.2.20.1 The interactor has typeCode *consumable* (CSM) for a substance being consumed (i.e., used up, disappearing in the process), or *catalyst* (CAT) for a catalytic substance being recovered and used again.

14.2.20.2 The interactor of the main process has type code CSM for the consumable source material, i.e., the source material.

14.2.20.3 The interactor may have one quantity as detailed in 14.2.19.3 and following.

14.2.20.4 If the interactor has a quantity, then the productOf element leading to the main derivation process also has a quantity.

14.2.20.5 The interactor has an identified substance.

14.2.20.6 The identified substance id has same extension and root as the nested identified substance code and codeSystem respectively.

14.2.20.7 The identified substance code is the UNII (2.16.840.1.113883.4.9).

14.2.20.8 If there is a name, then the name must match the UNII code.

14.2.21 Component Process

```
<derivationProcess>
  <interactor typeCode="CSM" .../>
  <component>
    <sequenceNumber value="1"/>
    <derivationProcess>
      <code code="C157220" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="antigen cultivation"/>
      <interactor typeCode="CSM" .../>
    </derivationProcess>
  </component>
```

```
  <component>
    <sequenceNumber value="2"/>
    <derivationProcess>
      <code code="C164239" codeSystem="2.16.840.1.113883.3.26.1.1"
            displayName="Ultraviolet Irradiation Attenuation" />
    </derivationProcess>
  </component>
</derivationProcess>
```

The terminology for the component derivation process codes (Derivation Process list) can be found at the URL:. ftp://ftp1.nci.nih.gov/pub/cacore/EVS/CBDD/CBDD_Process_Terminology.xls.

Validation Procedures

14.2.21.1 The derivationProcess has one or more component derivationProcesses.

14.2.21.2 A derivationProcess component has a positive integer sequence number, except if this is the only component.

14.2.21.3 Process component has a derivation process.

14.2.21.4 Derivation process has a code.

14.2.21.5 Code system is 2.16.840.1.113883.3.26.1.1

14.2.21.6 Code comes from the Derivation Process list (see text above).

14.2.21.7 Display name matches the code.

14.2.21.8 The component derivation process does not have nested component derivation process(es).

14.2.21.9 For agent modifications, the component derivation process may have one or more interactor(s) as in 14.2.20.

14.2.22 [RESERVED]